МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В.Ломоносова

Механико-математический факультет

Г.И.Фалин

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ДЛЯ ШКОЛЬНИКОВ

Москва
2017

УДК 519.2 (075.3)
ББК 22.172я729
Ф19

Фалин Г.И.

Элементы математической статистики для школьников. – М.: 2017 – 55 с., ил.

Мы рассказываем об основных понятиях статистики, понимая под ней, как это принято в школьном курсе, описательную статистику. Основное внимание уделено статистистическим переменным и операциям над ними (линейные преобразования, суммирование, смешивание). В заключительном разделе на примере истории открытия аргона мы рассказываем о роли статистики в естествознании.

Для учащихся старших классов и учителей математики.

Содержание

1 Основные понятия

1.1 Генеральная совокупность

В ходе любого статистического исследования изучается какое-то свойство (или несколько свойств) большой группы однотипных объектов. Эта группа называется *генеральной совокупностью* (от латинского generalis – главный).

Генеральная совокупность, как и всякое множество, может быть задана перечислением всех своих элементов или указанием общего свойства, которое объединяет объекты в эту совокупность. Если изучаемая группа состоит из небольшого числа объектов (например, школьный класс, в котором учится 20-40 учеников), то все эти объекты можно без всякого труда указать явно. При выставлении оценок по ЕГЭ по математике в 2011 году изучаемая группа содержала около 740 тысяч объектов (школьников), но использование современной компьютерной техники для сбора и хранения данных позволяет явно указать все объекты и в этом случае. В ряде случаев трудно даже подсчитать общее число объектов в генеральной совокупности (хотя совершенно ясно, какие объекты в неё входят). Например, при изучении рыбных запасов в озере Байкал генеральная совокупность включает всех рыб, живущих в этом озере. Ясно, что подсчитать их число точно практически невозможно (даже если не принимать в расчёт непрерывное изменение числа рыб). Тем не менее, существуют статистические методы, которые позволяют решить эту задачу с разумной точностью. С ещё более трудной проблемой сталкивается автомобильная компания, которая хочет понять, какие цвета предпочитают потенциальные покупатели новой модели автомобиля. В этом примере невозможно сразу точно сказать, какие люди входят в генеральную совокупность.

При статистическом исследовании чрезвычайно важно ясно понимать, о какой генеральной совокупности идёт речь, т.е. какая группа объектов исследуется. Без этого нельзя правильно организовать сбор и обработку данных, правильно интерпретировать полученные результаты и выработать рекомендации по их практическому применению.

1.2 Статистические переменные

Та характеристика объектов из генеральной совокупности, которая изучается в ходе статистического исследования, вообще говоря, меняется от объекта к объекту. Поэтому как свойство неопределённого объекта из рассматриваемой совокупности эта характеристика является *переменной (величиной)*. Поскольку эта переменная появляется в ходе статистического исследования, её называют *статистической переменной*. Мы для краткости будем опускать в дальнейшем слово «статистическая» и говорить просто о «переменной».

Статистические переменные принято обозначать каким-нибудь символом, обычно прописной (заглавной, большой) буквой латинского алфавита (*X*, *Y*, *Z* и т.д.) или строчной (маленькой) буквой греческого алфавита (ξ, ζ и т.д.). Точное *значение переменной* определяется тем, о каком конкретно объекте генеральной совокупности идёт речь.

Чтобы проиллюстрировать понятия генеральной совокупности и статистической переменной рассмотрим задачу 810 из учебника *Математика: Алгебра. Функции. Анализ данных. Учебник для 8 кл. общеобразоват. учреждений./Г.В.Дорофеев, С.Б.Суворова, Е.А.Бунимович и др; под ред. Г.В.Дорофеева. – М.: Просвещение, 2007.*:

«Маша, Саша, Катя, Лена, Ваня и Миша пошли в пиццерею. Ваня съел 5 кусков пиццы, Миша, Саша и Лена – по 3 куска, Катя – 2 куска, Маша – 1 кусок…»

В этом примере генеральная совокупность состоит из 6 «объектов» – школьников по именам Маша, Саша, Катя, Лена, Ваня, Миша. Для каждого школьника «измеряется» (фактически просто записывается) количество съеденных им кусков пиццы – это и будет рассматриваемой в задаче статистической переменной. Обозначая это количество буквой *W*, можно написать:

$$W(\text{Маша})=1,\ W(\text{Саша})=3,\ W(\text{Катя})=2,\ W(\text{Лена})=3,\ W(\text{Ваня})=5,\ W(\text{Миша})=3.$$

Если каждому школьнику присвоить номер (например, в соответствии с тем порядком, в котором их имена появились в условии задачи), то мы получим:

$$W(1)=1,\ W(2)=3,\ W(3)=2,\ W(4)=3,\ W(5)=5,\ W(6)=3.$$

Здесь числа в круглых скобках после буквы *W* – это объекты генеральной совокупности (точнее, их номера), а числа после знака равенства – *значения* изучаемой переменной.

Аргумент можно указывать и в виде нижнего индекса:

$$W_{\text{Маша}}=1,\ W_{\text{Саша}}=3,\ W_{\text{Катя}}=2,\ W_{\text{Лена}}=3,\ W_{\text{Ваня}}=5,\ W_{\text{Миша}}=3.$$

Соответственно, если объекты, входящие в генеральную совокупность, перенумерованы, то значения статистической переменной *W* можно обозначать $W_1, W_2, \ldots$ и т.п. и, скажем, в нашем примере писать:

$$W_1=1,\ W_2=3,\ W_3=2,\ W_4=3,\ W_5=5,\ W_6=3.$$

Ещё один пример. Предположим, что генеральная совокупность состоит из четырёх учеников, Пети, Маши, Саши, Даши, а интересующее нас свойство этих «объектов» – число решённых задач на вступительном экзамене в МГУ по математике (обычно на дополнительном вступительном испытании предлагают 8 задач).

Если, скажем, число решённых задач мы обозначаем буквой *X*, Петя решил 5 задач, Маша – 7, Саша – 3, Даша – 5, то этот факт можно выразить следующим образом:

$$X(\text{Петя})=5,\ X(\text{Маша})=7,\ X(\text{Саша})=3,\ X(\text{Даша})=5$$

или

$$X_{\text{Петя}} = 5,\ X_{\text{Маша}} = 7,\ X_{\text{Саша}} = 3,\ X_{\text{Даша}} = 5.$$

1.3 Числовые переменные и наборы

Если значения статистической переменной являются числами, её называют *числовой* (или *количественной*) переменной.

В математической статистике содержательный смысл значений числовых переменных, как правило, не играет никакой роли; обычно это просто *наборы чисел*. Соответственно, *набор значений числовой переменной* мы назовём *числовым набором*. Скажем, в нашем примере с ДВИ по математике набор имеет вид: 5, 7, 3, 5. Здесь и позже мы будем разделять элементы набора друг от друга запятой, а для записи десятичных дробей вместо десятичной запятой будем использовать точку, так что, например, $1/2 = 0.5$.

В статистике обычно работают только с набором значений переменной и фактически отождествляют статистическую переменную и набор её значений. Но важно понимать, что числовые наборы в статистике – это не какие-то произвольные, неизвестно откуда появившиеся, группы чисел. Любой числовой набор в статистике всегда является результатом измерения какого-то количественного свойства объектов определённой генеральной совокупности, а каждое число из набора относится к *одному* объекту этой генеральной совокупности. Эти замечания относятся и к нечисловым переменным, о которых мы будем говорить позже.

Школьник, знакомый с основами теории множеств в объёме школьного курса, без труда поймёт, что статистическая числовая переменная – это отображение множества изучаемых объектов (генеральной совокупности) в множество действительных чисел.

Следует отметить, что в специальных разделах статистики (медицинская статистика, финансовая статистика и т.д.) для получения практически значимых результатов необходимо в определённой мере принимать в расчёт соответствующую (медицинскую, финансовую и т.д.) информацию, содержащуюся в числах, образующих числовой набор.

Понятие «набор» нуждается в дополнительных комментариях. В рассматриваемом нами примере для статистического исследования уровня знаний по математике не играет роли, кто конкретно из ребят решил 7 задач, а кто 3. Если бы Маша решила 3 задачи, а Саша – 7, то это никак не повлияло бы на общую оценку уровня подготовки нашей группы школьников (хотя, несомненно, обрадовало бы родителей Саши и очень расстроило бы родителей Маши). С этой точки зрения наборы [5,7,3,5] и [5,3,7,5] следует считать одинаковыми. Поэтому мы уточним определение числового набора следующим образом:

Числовой (статистический) набор – *это конечная неупорядоченная последовательность чисел.*

Несложно дать и аккуратное математическое определение числового статистического набора, например, как класса последовательностей одинаковой длины, отличающихся друг от друга перестановкой элементов (считая тем самым последовательности одинаковой длины, отличающиеся друг от друга перестановкой элементов, «эквивалентными», т.е. реализациями одного и того же неупорядоченного набора).

Чтобы не усложнять обозначения, мы будем обозначать набор значений переменной тем же символом, что и саму переменную (как правило, заглавной буквой латинского алфавита), а значения переменной для индивидуальных объектов из генеральной совокупности – соответствующей прописной буквой с нижним индексом, указывающим на номер объекта в генеральной совокупности при некоторой их нумерации. Таким образом, в рассматриваемом примере можно писать: набор $X = [x_1,\ x_2,\ x_3,\ x_4]$, где $x_1 = 5,\ x_2 = 7,\ x_3 = 3,\ x_4 = 5$.

В математике запись вида {5,7} обычно означает *множество* из элементов 5 и 7, а запись (5,7) – *упорядоченную* последовательность из двух элементов 5 и 7. Кроме того, математики считают, что

- {5,7}={7,5}, т.к. элементы множества не упорядочены,
- запись {5,7,3,5} не имеет смысла, т.к. в множестве каждый элемент указывается только один раз (с некоторой натяжкой её можно было бы отождествить с записью {5,7,3}),
- $(5,7,3,5) \neq (3,5,5,7)$, т.к. последовательности, отличающиеся порядком элементов, считаются различными.

По этой причине мы решили употреблять для обозначения *неупорядоченного* набора [5,7,3,5] квадратные скобки. Соответственно, мы считаем, что $[5,7,3,5] = [3,5,5,7]$.

Отметим, что в некоторых статистических исследованиях порядок, в котором расположены анализируемые значения, очень важен. В этих случаях мы будем указывать числа, образующие набор, в круглых скобках. Рассмотрим, например, Таблицу 1.1, в которой указана средняя температура в некотором регионе в июле за последние 8 лет:

Таблица 1.1

год	2009	2010	2011	2012	2013	2014	2015	2016
температура	21.2°	22.1°	23.2°	22.6°	22.9°	23.6°	24.8°	24.6°

Если мы хотим подсчитать среднюю температуру за 8 лет, то порядок чисел в наборе [21.2, 22.1, 23.2, 22.6, 22.9, 23.6, 24.8, 24,6] не играет никакой роли. Но если мы хотим понять, происходит ли изменение климата, то порядок очень важен (в данном регионе явно видна тенденция роста среднемесячной температуры в июле) и чтобы подчеркнуть, что рассматривается упорядоченный набор, мы будем записывать его в круглых скобках: (21.2, 22.1, 23.2, 22.6, 22.9, 23.6, 24.8, 24.6).

Основными характеристиками числовой переменной (числового набора) являются среднее значение

$$\mathrm{M}(X) = \frac{x_1 + \ldots + x_n}{n},$$

дисперсия

$$\mathrm{D}(X) = \frac{\left[x_1 - \mathrm{M}(X)\right]^2 + \ldots + \left[x_n - \mathrm{M}(X)\right]^2}{n}$$

и стандартное отклонение $\sigma(X) = \sqrt{\mathrm{D}(X)}$.

Формулу, определяющую дисперсию, можно переписать в более удобной форме, если проделать следующие преобразования:

$$\mathrm{D}(X)=\frac{\left[x_1^2-2\mathrm{M}(X)x_1+\left(\mathrm{M}(X)\right)^2\right]+\ldots+\left[x_n^2-2\mathrm{M}(X)x_n+\left(\mathrm{M}(X)\right)^2\right]}{n}$$

$$=\frac{x_1^2+\ldots+x_n^2}{n}-2\mathrm{M}(X)\cdot\frac{x_1+\ldots+x_n}{n}+\frac{n\cdot\left(\mathrm{M}(X)\right)^2}{n}$$

$$=\frac{x_1^2+\ldots+x_n^2}{n}-2\mathrm{M}(X)\cdot\mathrm{M}(X)+\left(\mathrm{M}(X)\right)^2=\frac{x_1^2+\ldots+x_n^2}{n}-\left(\mathrm{M}(X)\right)^2.$$

Обозначим набор $\left[x_1^2,\ldots,x_n^2\right]$ через X^2 (подробнее об операциях над числовыми наборами или, что то же самое, над статистическими переменными мы будем говорить позже). Тогда $\frac{x_1^2+\ldots+x_n^2}{n}$ можно рассматривать как среднее значение набора X^2. Это даёт следующую формулу для дисперсии:

$$\mathrm{D}(X)=\mathrm{M}\left(X^2\right)-\left(\mathrm{M}(X)\right)^2.$$

В заключение ещё раз подчеркнём, что любой числовой набор $X=[x_1,\ldots,\,x_n]$ в статистике появляется как результат измерения определённого свойства, характеристики и т.п. объектов из некоторой генеральной совокупности, т.е., в нашей терминологии, как набор значений определённой статистической переменной. Число n элементов набора равно числу объектов в рассматриваемой генеральной совокупности. Поэтому когда мы говорим о среднем значении, дисперсии, функции распределения и т.д. числового набора мы фактически говорим о среднем значении, дисперсии, функции распределения и т.д. этой переменной.

1.4 Дискретные числовые переменные

Числовые переменные в статистике, в свою очередь, делятся на два типа: *дискретные* и *непрерывные*.

Числовая переменная называется ***дискретной*** (от латинского discretus – разделённый, прерывистый)*, если её значения ясно отделены друг от друга.*

Набор значений дискретной числовой переменной $X=[x_1,\ldots,\,x_n]$, как правило, содержит много повторяющихся значений. Иначе говоря, количество N *различных* чисел в наборе $[x_1,\ldots,\,x_n]$ обычно намного меньше числа n (которое равно числу объектов в рассматриваемой генеральной совокупности). Поэтому удобно рассматривать не набор $X=[x_1,\ldots,x_n]$, а

1. ввести набор $y_1,\ldots,y_N$ различных значений переменной X,
2. для каждого значения y из списка $y_1,\ldots,y_N$ подсчитать число $t(y)$ его повторений в исходном наборе,

после чего описать набор $X=[x_1,\ldots,\,x_n]$ более компактным набором пар $\left(y_k,t\left(y_k\right)\right)$, $k=1,\ldots,N$.

Обычно удобно включать в набор $y_1,\ldots,y_N$ не просто *различные значения* переменной X, а все *теоретически возможные её значения* (исходя из

содержательного смысла этой переменной). Например, если генеральная совокупность – это все дни декабря 2016 года, переменная X указывает число ДТП в Москве в конкретный день, то её теоретически возможные значения – это все неотрицательные целые числа. Конечно, с точки зрения здравого смысла считать, что X может равняться, например, 100 тысяч совершенно нелепо. Но в равной степени нелепо считать, что X может быть равно n, но не может быть равно n+1. При таком, более широком, взгляде на понятие «возможное значение» нельзя исключить, что некоторое значение y, хотя и является теоретически возможным значением анализируемой переменной X, на самом деле в наборе $[x_1,\ldots,x_n]$ не встречается. В этом случае $t(y)=0$.

Отметим очевидное соотношение, которое часто будет использоваться позже:

$$t(y_1)+\ldots+t(y_n)=1. \qquad (1.1)$$

Как правило, подсчитывают и (относительную) частоту повторения каждого значения $f(y_k)=t(y_k)/n$. Для частот соотношение (1.1) превратится в равенство:

$$f(y_1)+\ldots+f(y_n)=1. \qquad (1.2)$$

Результаты этой первичной статистической обработки представляют в табличной или графической (всевозможные диаграммы) форме. Например, для статистического исследования результаты ДВИ по математике в группе из $n=50$ абитуриентов могут быть оформлены в виде Таблицы 1.2 (мы считаем, что вариант содержит 8 задач):

Таблица 1.2

число решённых задач, y	0	1	2	3	4	5	6	7	8
число абитуриентов, решивших данное число задач, $t(y)$	4	2	1	8	12	11	8	3	1
доля, $f(y)=t(y)/n$	8%	4%	2%	16%	24%	22%	16%	6%	2%

Столбиковая диаграмма, соответствующая этой таблице, изображена на Рисунке 1.1; она даёт общее представление о характере *распределения* числа решённых задач.

Рисунок 1.1

На этой диаграмме для каждого возможного значения y рассматриваемой переменной мы отметили абсолютное число повторений этого значения (обозначенное выше через $t(y)$). С равным успехом можно было бы указывать (относительную) частоту $f(y) = t(y)/n$ повторения каждого значения. Соответствующая диаграмма отличается от уже построенной только масштабом по оси ординат.

Набор относительных частот $f(y_1),\ldots,f(y_N)$, который указывает, насколько часто переменная принимает теоретически возможные значения, называется *распределением* этой переменной. Если нужно явно указать, к какой переменной относится распределение, мы будем добавлять соответствующий нижний индекс, так что, например, $f_X(y)$ – распределение переменной X, $f_Y(y)$ – распределение переменной Y и т.д.

Для дискретных переменных столбики отделены друг от друга небольшим промежутком. Другие виды диаграмм (особенно популярны круговые диаграммы) обычно дают только приблизительное визуальное впечатление о характере распределения и чаще всего используются только для *украшения* статьи, доклада и пр.

Столбиковые диаграммы используются и для графического описания характера распределения нечисловых переменных (об этом виде переменных мы будем говорить ниже).

С помощью распределения числовой переменной удобно вычислять её основные характеристики – среднее значение и дисперсию.

Пусть, как мы определили выше, N – количество возможных значений переменной $X = \left[x_1,\ldots,x_n\right]$, а $y_1,\ldots,y_N$ – сами эти значения. Обычно мы будем считать, что значения $y_1,\ldots,y_N$ занумерованы в порядке возрастания, но сейчас это несущественно. Пусть, далее, натуральное число $t_k = t\left(y_k\right)$ указывает, сколько раз число y_k встречается в наборе. Группируя вместе одинаковые числа

основного набора $X=\left[x_1,\ldots,x_n\right]$, мы получим, что сумма $x_1+\ldots+x_n$ равна $t(y_1)\cdot y_1+\ldots+t(y_N)\cdot y_N$. Соответственно, среднее значение $\mathrm{M}(X)$ равно:

$$\begin{aligned}\mathrm{M}(X)&=\frac{t(y_1)y_1+\ldots+t(y_N)y_N}{n}=y_1\frac{t(y_1)}{n}+\ldots+y_N\frac{t(y_N)}{n}\\&=y_1f_X(y_1)+\ldots+y_Nf_X(y_N).\end{aligned}\qquad(1.3)$$

Для дисперсии аналогично имеем:

$$\begin{aligned}\mathrm{D}(X)&=\frac{t(y_1)\cdot\left[y_1-\mathrm{M}(X)\right]^2+\ldots+t(y_N)\cdot\left[y_N-\mathrm{M}(X)\right]^2}{n}\\&=\left[y_1-\mathrm{M}(X)\right]^2\cdot\frac{t(y_1)}{n}+\ldots+\left[y_N-\mathrm{M}(X)\theta\right]^2\cdot\frac{t(y_N)}{n}\\&=\left[y_1-\mathrm{M}(X)\right]^2\cdot f_X(y_1)+\ldots+\left[y_N-\mathrm{M}(X)\right]^2\cdot f_X(y_N).\end{aligned}\qquad(1.4)$$

Формулы (1.3) и (1.4) удобны для расчёта среднего значения и дисперсии в случае, когда данные представлены в сгруппированной форме и известно распределение $f_X(y_1),\ldots,f_X(y_N)$ значений анализируемой переменной X.

1.5 Непрерывные числовые переменные

*Числовая переменная называется **непрерывной**, если её значением может быть любое действительное число из некоторого промежутка (по крайней мере, теоретически).*

Непрерывными являются все переменные, которые выражают физические характеристики объекта: длину, массу, объём, температуру, время и т.д.

При физических измерениях мы всегда определяем значения длины, массы и т.д. приближённо, с точностью, зависящей от вида используемого прибора, так что результат всегда выражается десятичной дробью с конечным числом знаков после запятой. При подходящем выборе единицы измерения результат будет выражается целым числом. Примерно так же обстоит дело и с денежными суммами – они всегда выражаются целым числом копеек. Поэтому с практической точки зрения непрерывных переменных нет вовсе. Однако теоретически измеряемые величины могут выражаться любым положительным действительным числом.

Важно учесть и ещё одно соображение. Дискретные переменные характеризуются тем, что их значения ясно отделены друг от друга. Но разница между 456 руб. 89 коп. и 457 руб. 23 коп. настолько мала, что с практической точки зрения вряд ли разумно считать эти суммы различными. Если денежные суммы отмечать точками на числовой оси, то даже при выборе единицы масштаба в 1 руб. разница между соседними точками (величиной 0.01) будет настолько мала, что эти дискретные точки сольются в непрерывную линию. Это соображение показывает, что в ряде случаев формально дискретные переменные иногда удобно считать непрерывными.

Для непрерывных переменных любое конкретное значение малоинтересно. Например, при анализе доходов населения совершенно неважно, сколько человек имели в прошлом году суммарный доход в размере 327851 руб. 37 коп. Действительно интересно, например, сколько человек имели годовой доход меньше, чем 150 тыс. руб., т.е. количество малообеспеченных, сколько человек имели годовой доход от 600 тыс. руб. до 1 млн. руб., т.е. количество достаточно обеспеченных, и т.д. Таким образом, для непрерывных переменных разумно обращать внимание лишь на классы значений, попадающих в тот или иной промежуток. Поэтому первый шаг в статистической обработке непрерывных переменных заключается в разбиении диапазона возможных значений на несколько промежутков (их называют интервалами группировки) и вычислении числа значений, попадающих в каждый интервал. Интервалы группировки должны быть *непересекающимися* (нельзя, чтобы какое-то значение можно было бы отнести к двум разным интервалам) и *исчерпывающими* (т.е. каждое теоретически возможное значение должно быть отнесено в какой-то интервал).

По поводу интервалов группировки необходимо сделать важное замечание, связанное с процедурой округления значений непрерывных переменных. Предположим, например, что ученик на вопрос, сколько времени он сегодня потратил на дорогу до школы, ответил: «18 минут». Если время округлялось до целых минут отбрасыванием секунд, то значение 18 означает, что точное время T лежит в промежутке $18 \le T < 19$. Если же время округлялось до целых минут по обычному правилу (до ближайшего целого), то значение 18 означает, что точное значение времени на дорогу до школы удовлетворяет двойному неравенству $17.5 \le T < 18.5$.

Если мы фиксируем время на дорогу для школы для большой группы учеников и собираем в один класс все значения переменной T от 10 до 19 включительно, то при первом способе округления этот класс характеризуется неравенством $10 \le T < 20$, а при втором – неравенством $9.5 \le T < 19.5$. В каждом случае длина интервала группировки равна 10, но сами интервалы разные. В частности, в первом случае центр интервала (он важен для расчёта статистических характеристик по сгруппированным значениям) равен 15, а во втором этим центром будет точка 14.5. Чтобы избежать недоразумений, лучше точно описывать классы группировки двойными неравенствами. В школьных учебниках на процедуру округления числовых данных вообще не обращают внимания, что недопустимо при аккуратном статистическом анализе непрерывных переменных.

Общее визуальное представление о характере распределения значений непрерывной переменной даёт гистограмма. Мы кратко опишем процедуру её построения на примере следующего набора T из 100 чисел:

27, 52, 43, 38, 47, 8, 21, 40, 32, 53, 45, 54, 35, 28, 40, 18, 31, 45, 24, 30,
37, 15, 39, 34, 48, 25, 30, 7, 32, 12, 26, 35, 48, 19, 33, 26, 17, 30, 42, 22,
53, 28, 42, 36, 23, 10, 34, 46, 16, 29, 35, 52, 41, 32, 21, 39, 55, 25, 29, 8,
36, 44, 26, 55, 34, 19, 42, 54, 27, 10, 45, 20, 31, 50, 18, 9, 41, 14, 38, 40,
23, 49, 33, 15, 24, 46, 36, 28, 32, 37, 51, 20, 29, 47, 33, 27, 41, 22, 39, 40.

Этот набор мы взяли из учебника *Математика: алгебра. Функции. Анализ данных. Учебник для 9 кл. общеобразоват. учреждений.*/Г.В.Дорофеев, С.В.Суворов, Е.А.Бунимовач и др; под ред. Г.В.Дорофеева. – М.: Просвещение,

2005., стр. 251. Числа набора – это время (в минутах), которое 100 учеников гипотетической школы тратят на дорогу в школу.

К сожалению, мы не знаем, как собиралась эти данные – кто-то уже собрал их за нас (в статистике такие данные называют *вторичными*). В частности, мы не знаем использовавшуюся процедуру округления. Для определённости будем считать, что время округлялось до целых минут отбрасыванием секунд. Поэтому, например, в промежуток $20 \le T < 25$ мы будем включать значения 20, 21, 22, 23, 24 (в рассматриваемом наборе 10 таких значений).

Беглый взгляд на исходные данные показывает, что возможные значения переменной T удовлетворяют не превосходят 60 минут, причём большинство школьников тратит на дорогу от 20 до 50 минут. Этот промежуток мы разобьём на 6 равных интервалов группировки, длиной 5 минут каждый. В каждый из них попадает около 10 чисел набора. В интервалы $0 < T < 20$ и $50 < T < 60$ попадает 16 и 10 чисел соответственно. Поэтому мы не будем дробить их на более мелкие промежутки. В результате мы получим 8 интервалов группировки:

$$0 < T < 20,\ 20 \le T < 25,\ 25 \le T < 30,\ 30 \le T < 35,\ 35 \le T < 40,\ 40 \le T < 45,\ 45 \le T < 50,\ 50 \le T < 60.$$

Длины этих интервалов в минутах равны: 20, 5, 5, 5, 5, 5, 5, 10 соответственно (обратим внимание на то, что специфика задачи потребовала рассматривать интервалы группировки разной длины) и они содержат 16, 10, 14, 15, 13, 12, 10, 10 значений исходного набора T соответственно. В отличие от дискретных переменных, для непрерывных переменных принято рассматривать не обычные (относительные) частоты $f_i = \frac{n_i}{n}$ попадания в i-й класс (здесь n_i – количество значений в i-м классе), а *нормированные* частоты $f_i^* = \frac{n_i}{n \cdot \Delta_i}$, где Δ_i – длина i-го интервала группировки. В нашем случае эти величины приведены в Таблице 1.3.

Таблица 1.3

Интервал группировки	Длина интервала, Δ_i (мин)	Количество чисел в классе, n_i	Нормированная частота, $f_i^* = \frac{n_i}{n\Delta_i}$
$0 < T < 20$	20	16	0.008
$20 \le T < 25$	5	10	0.020
$25 \le T < 30$	5	14	0.028
$30 \le T < 35$	5	15	0.030
$35 \le T < 40$	5	13	0.026
$40 \le T < 45$	5	12	0.024
$45 \le T < 50$	5	10	0.020
$50 \le T < 60$	10	10	0.010
Всего	**60**	**100**	

Теперь можем постоить гистограмму. По определению, гистограмма – это функция $h(x)$ действительного аргумента x, которая на i-м промежутке группировки принимает постоянное значение, равное нормированной частоте $f_i^* = \dfrac{n_i}{n \cdot \Delta_i}$. Удобно также считать, что вне промежутка возможных значений анализируемой переменной гистограмма тождественно равна 0.

Для рассматриваемого набора T гистограмма изображена на Рисунке 1.2. Гистограмма очень похожа на столбиковую диаграмму для дискретной переменной, но (в отличие от столбиковой диаграммы) на гистограмме столбики стоят вплотную друг к другу и, вообще говоря, имеют разную ширину.

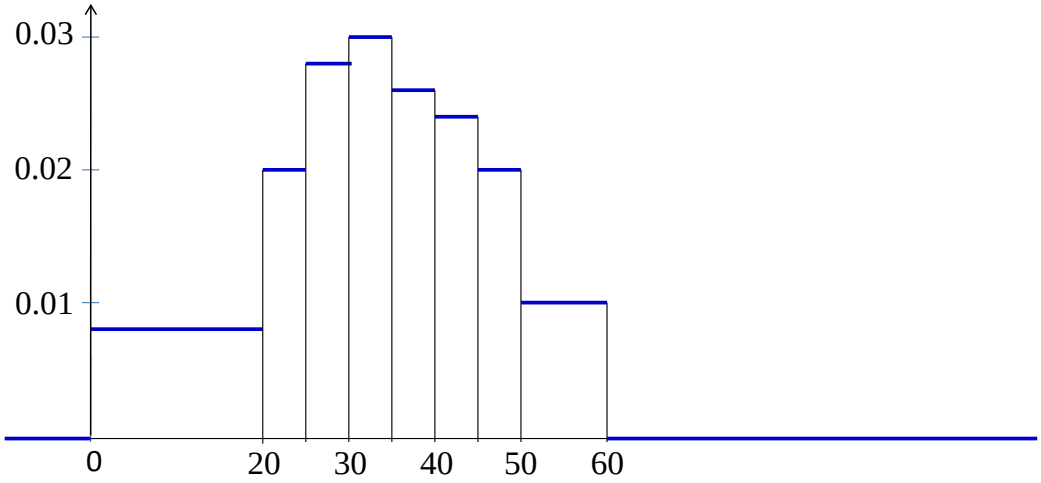

Рисунок 1.2

1.6 Качественные переменные и наборы

Предположим теперь, что интересующее нас свойство ученика из группы {Петя, Маша, Саша, Даша} – не число решённых задач на ДВИ по математике в МГУ, а его любимый предмет. Как и число решённых задач, вообще говоря, этот предмет меняется от ученика к ученику. Поэтому любимый предмет ученика как свойство неопределённого ученика из рассматриваемого класса является *переменной величиной*. Подобно тому как мы обозначили число решённых задач буквой X, эту переменную можно обозначить каким-нибудь символом, например, F. Точное *значение* этой *переменной* определяется тем, о каком конкретно из учащихся идёт речь. Если, например, Петя больше всего любит физику, Маша – математику, Саша – историю, Даша – химию, то этот факт можно выразить, например, следующим образом:

$$F_{\text{Петя}} = \text{физика}\,,\ F_{\text{Маша}} = \text{математика}\,,\ F_{\text{Саша}} = \text{история}\,,\ F_{\text{Даша}} = \text{химия}\,.$$

Для статистического исследования интересов учащихся не играет роли, кто из учеников любит какой предмет. Важно знать лишь общий *набор* значений

рассматриваемой переменной, т.е. любимых предметов школьников. Этим набором является неупорядоченная последовательность из четырёх слов: [физика, математика, история, химия].

Поскольку значения нашей переменной F не являются числами, а описывают общее качество рассматриваемого объекта, её называют *нечисловой* или *качественной* переменной, а соответствующий набор – *нечисловым* набором.

Неформально, можно сказать, что

Нечисловой (статистический набор) *– это просто конечная неупорядоченная последовательность слов, букв и т.д.*

Определение значения качественной переменной обычно заключается в анализе для каждого объекта свойства, которое описывает эта переменная, и отнесении этого объекта к одному из нескольких *классов* (в статистике эти классы называют *категориями*). Классы должны быть *взаимно исключающими* (ни один объект не может быть отнесён к двум классам) и *исчерпывающими* (классы должны покрывать все возможности). Этот процесс часто субъективен и потому по сравнению с количественными переменными статистический анализ качественных переменных труднее, а выводы менее надёжны.

Часто для удобства сбора и обработки статистических данных значения качественных переменных кодируют с помощью чисел. Мы могли бы, например, присвоить математике код 1, физике код 2, истории код 3, химии код 4. Тогда набор [физика, математика, история, химия] принял бы вид: [2,1,3,4]. Однако мы не будем называть его числовым, т.к. обычные действия над числами (сравнение, сложение и т.д.) в этой ситуации не имели бы никакого смысла. Чтобы подчеркнуть это обстоятельство, значения такой переменной F и саму переменную F называют *номинальными* (от латинского nomen – имя).

Обычные школьные оценки «2», «3», «4», «5» в сущности являются кодами для качественных оценок «неудовлетворительно», «удовлетворительно», «хорошо», «отлично». Эти нечисловые оценки мы можем совершенно точно *линейно упорядочить*: «отлично» лучше, чем «хорошо», «хорошо» лучше, чем «удовлетворительно», «удовлетворительно» лучше, чем «неудовлетворительно». Такие переменные называют *порядковыми* или *ординальными* (от латинского ordinatus – расположенный в порядке). Числовые коды оценок соответствуют этому порядку. Поэтому числовые оценки лучше считать качественными ординальными. Рассматривать их как количественные в строгом смысле этого слова можно только с определённой натяжкой: разница между «3» и «2» вовсе не равнозначна разнице между «5» и «4» и её вообще нельзя измерить числом. Суммирование оценок при определении «средней» оценки или «общего» балла можно оправдать только простотой процедуры. Особенно нелепо это при суммировании оценок по разным экзаменам, что приводит, например, к выводу, что 90 баллов по ЕГЭ и 50 баллов по независимому экзамену, проводимому МГУ (сумма баллов равна 140), лучше, чем 60 баллов по ЕГЭ и 75 баллов по независимому экзамену, проводимому МГУ (сумма баллов равна 135). Несмотря на это школьные оценки традиционно считаются числовыми и, чтобы упростить изложение, мы будем следовать этой традиции.

При статистическом исследовании чрезвычайно важно ясно понимать тип анализируемой переменной, так как от этого зависит сам характер этого исследования: как представлять данные, какие характеристики вычислять и т.д.

1.7 Вариационный ряд

Если элементы числового статистического набора упорядочить по возрастанию, то получившаяся версия набора является последовательностью в обычном смысле и называется *вариационным рядом*. Например, для набора [5,7,3,5] вариационный ряд имеет вид: (3,5,5,7).

В математической статистике k-й член вариационного ряда набора $X=\left[x_1,x_2,\ldots,x_n\right]$ обычно обозначают $x_{(k)}$; он является определённой функцией от всего этого набора. В частности, $x_{(1)}=\min\left(x_1,x_2,\ldots,x_n\right)$ – наименьшее число набора, $x_{(n)}=\max\left(x_1,x_2,\ldots,x_n\right)$ – наибольшее число набора.

Если мы сразу будем рассматривать не набор, а его вариационный ряд, то для его записи мы будем использовать обычное обозначение для последовательности (с круглыми скобками): $X=\left(x_{(1)},x_{(2)},\ldots,x_{(n)}\right)$.

Хотя данное выше определение вариационного ряда применимо и к качественному ординальному набору, содержательная теория существует только для числовых наборов.

С помощью вариационного ряда определяется ещё одна мера положения числового набора $X=\left[x_1,\ldots,x_n\right]$ – медиана $\mu(X)$. По определению, при нечётном числе элементов набора (когда $n=2k-1$ для некоторого натурального k) медиана равна k-му члену вариационного ряда; если же набор состоит из чётного числа элементов (т.е. $n=2k-1$ для некоторого натурального k), то медиана – это среднее арифметическое k-го и $(k+1)$-го членов вариационного ряда:

$$\mu(X)=\begin{cases} x_{(k)}, & \text{если } n=2k-1,\\ \dfrac{x_{(k)}+x_{(k+1)}}{2}, & \text{если } n=2k. \end{cases}$$

Медиане как мере положения в качестве меры разброса значений статистической переменной X соответствует размах, который определяется как разница между её наибольшим и наименьшим значениями:

$$\mathrm{R}(X)=x_{(n)}-x_{(1)}\equiv\max\left(x_1,\ldots,x_n\right)-\min\left(x_1,\ldots,x_n\right).$$

Размах равен длине отрезка, на котором расположены все числа набора.

1.8 Двумерные переменные

Часто при статистическом исследовании генеральной совокупности для составляющих её объектов измеряют не одно, а несколько свойств. Предположим, например, что для изучения уровня математической подготовки группы из 5 школьников им предложили две контрольные работы – одну по алгебре, а вторую по геометрии. Оценки записали в Таблицу 1.4.

Таблица 1.4

Номер	Имя	Алгебра	Геометрия
1	Петя	5	4
2	Саша	4	4
3	Коля	4	3
4	Витя	5	3
5	Дима	4	5

В этом случае мы имеем дело с двумя статистическими переменными: A (оценка по алгебре) и G (оценка по геометрии). Если эти переменные изучаются изолированно друг от друга, то каждую из этих переменных можно отождествить с набором её значений: A=[5,4,4,5,4], G=[4,4,3,3,5]. При этом для удобства статистического исследования вполне можно перейти к вариационным рядам, т.е. считать, что $A=\left(4,4,4,5,5\right)$, $G=\left(3,3,4,4,5\right)$. Однако, если мы хотим понять, имеется или нет зависимость между успеваемостью по алгебре и успеваемостью по геометрии, то необходимо для каждого школьника i соответствующие значения этих переменных, a_i и g_i, рассматривать как *неразрывную* пару (a_i, g_i). Иначе говоря, необходимо рассматривать *двумерную* переменную (A,G). Набор её значений – это неупорядоченный набор из упорядоченных пар: [(5,4), (4,4), (4,3), (5,3), (4,5)].

По аналогии с двумерными можно рассматривать *трёхмерные*, *четырёхмерные* и, вообще, *многомерные* переменные.

Ещё раз подчеркнём, что каждое значение (x_i, y_i) двумерной переменной (X,Y) относится к *одному* объекту какой-то генеральной совокупности. Если переменные X и Y описывают объекты двух *разных* генеральных совокупностей, состоящих из одинакового числа n объектов, то можно формально образовать из наборов $[x_1,\ldots,x_n]$ и $[y_1,\ldots,y_n]$ значений этих переменных набор $[(x_1,y_1),\ldots, (x_n,y_n)]$ из n пар (x_i,y_i) (при этом имеется много возможностей скомпоновать пары). *Но этот набор не будет иметь никакого содержательного смысла.* В частности, к нему нельзя применять теорию статистической зависимости.

2 Линейные преобразования статистических переменных

2.1 Определение линейного преобразования

Пусть X – некоторая числовая статистическая переменная, заданная на генеральной совокупности из n объектов. Мы будем считать, что эти объекты как-то перенумерованы и обозначим через $X(i)$ значение нашей переменной для i-го объекта.

Если для каждого объекта i число $X(i)$ умножить на одно и то же число a, то мы получим новую переменную X', значения которой связаны со значениями переменной X формулой: $X'(i) = aX(i)$, $i = 1,\ldots,n$. Эту переменную мы будем обозначать aX (или Xa) и называть произведением числа a и переменной X.

Если $a \neq 0$, то можно определить операцию деления переменной X на число a как умножение X на число $1/a$: $\dfrac{X}{a} \equiv \dfrac{1}{a} \cdot X$.

Аналогично, если для каждого объекта i к числу $X(i)$ прибавить одно и то же число b, то мы получим новую переменную X'', значения которой связаны со значениями переменной X формулой: $X''(i) = X(i) + b$, $i = 1,\ldots,n$. Эту переменную мы будем обозначать $X + b$ (или $b + X$) и называть суммой переменной X и числа b.

Естественно определить и операцию вычитания из переменной X числа b как сумму X и $-b$: $X - b \equiv X + (-b)$.

Можно объединить обе операции и для каждого объекта $i = 1,\ldots,n$ генеральной совокупности

1. сначала умножить число $X(i)$ на одно и то же число a,
2. затем к полученному числу $aX(i)$ прибавить одно и то же число b, что даст число $aX(i) + b$.

В результате мы получим новую переменную Y, значения которой связаны со значениями переменной X формулой: $Y(i) = aX(i) + b$, $i = 1,\ldots,n$. Эту переменную естественно обозначить $aX + b$. Замена исходной переменной X на переменную $Y = aX + b$ называется *линейным преобразованием* (по аналогии с обычной линейной функцией, которая задаётся формулой $y = ax + b$).

Если статистическую переменную X отождествить с набором $[x_1,\ldots,x_n]$ её значений, то мы можно говорить о *линейном преобразовании* $Y = aX + b$ *числового набора*. Под этим термином мы понимаем замену набора $X = [x_1,\ldots,x_n]$ на набор $Y = [ax_1 + b,\ldots,ax_n + b]$, где a и b – некоторые константы, и потому часто вместо записи $Y = aX + b$ мы будем использовать запись $y_i = ax_i + b$ или, короче, $y = ax + b$ (т.е. указывать ту линейную функцию, которая определяет значения y_i набора Y по значениям x_i набора X).

Параметр a называется *масштабным коэффициентом* (или *масштабным параметром*; используют также и термин «*нормирующий параметр*»), а параметр b – *параметром сдвига* (или, короче, *сдвигом*; используют также и термин «*центрирующий параметр*»).

Пример 2.1. Если к набору $X=\left[1.25,\ 1.21,\ 1.19\right]$ применить преобразование $Y=100X-120$, то мы получим набор $Y=\left[5,\ 1,-1\right]$. В этом примере масштабный коэффициент $a=100$, а параметр сдвига $b=-120$.

Чтобы в будущем исключить случаи, которые не представляют интереса для приложений, а лишь усложняют формулировки утверждений, мы будем рассматривать только линейные преобразования с ненулевым коэффициентом a.

Линейные преобразования числовых статистических переменных или, что то же самое, линейные преобразования наборов их значений обладают несколькими важными свойствами.

Свойство 1 (*рефлексивность*) Любую числовую статистическую переменную X можно получить из самой себя линейным преобразованием $aX+b$ с коэффициентами $a=1$, $b=0$. Иначе говоря, любой числовой набор $X=\left[x_1,\ldots,x_n\right]$ можно получить из самого себя линейным преобразованием $aX+b$ с коэффициентами $a=1$, $b=0$.

Свойство 2 (*симметричность*) Если числовая статистическая переменная Y получена из переменной X линейным преобразованием $Y=aX+b$, то переменная X может быть получена из переменной Y линейным преобразованием $X=\frac{1}{a}\cdot Y+\left(-\frac{b}{a}\right)$. Иначе говоря, если набор $Y=\left[y_1,\ldots,y_n\right]$ получен из набора $X=\left[x_1,\ldots,x_n\right]$ линейным преобразованием $Y=aX+b$, то набор $X=\left[x_1,\ldots,x_n\right]$ может быть получен из набора $Y=\left[y_1,\ldots,y_n\right]$ линейным преобразованием $X=\frac{1}{a}\cdot Y+\left(-\frac{b}{a}\right)$; его можно представить и формулой $X=\frac{Y-b}{a}$.

Обратим внимание на то, что в последнем равенстве коэффициент $\frac{1}{a}$ не равен нулю.

Свойство 3 (*транзитивность*) Если числовая статистическая переменная Y получена из переменной X линейным преобразованием $Y=a'X+b'$, а переменная Z, в свою очередь, получена из переменной Y линейным преобразованием $Z=a''Y+b''$, то переменная Z может быть получена из переменной X линейным преобразованием $Z=a'a''X+\left(a''b'+b''\right)$. Иначе говоря, если набор $Y=\left[y_1,\ldots,y_n\right]$ получен из набора $X=\left[x_1,\ldots,x_n\right]$ линейным преобразованием $Y=a'X+b'$, а набор $Z=\left[z_1,\ldots,z_n\right]$ – из набора $Y=\left[y_1,\ldots,y_n\right]$ линейным преобразованием $Z=a''Y+b''$, то набор $Z=\left[z_1,\ldots,z_n\right]$ может быть получен из набора $X=\left[x_1,\ldots,x_n\right]$ линейным преобразованием $Z=a'a''X+\left(a''b'+b''\right)$.

Обратим внимание на то, что коэффициент $a'a''$ не равен нулю (так как отличны от нуля коэффициенты a' и a'').

В высшей алгебре связь между парами объектов (любой природы), обладающую свойствами рефлексивности, симметричности и транзитивности называют *отношением эквивалентности*. Эквивалентные объекты в некотором смысле можно считать неотличимыми друг от друга (или различными представлениями одного и того же объекта). В нашем случае наборы, отличающиеся друг от друга линейным преобразованием, в сущности задают один и тот же набор данных, но измерены эти данные в разных единицах.

Значение линейных преобразований в описательной статистике заключается в том, что они упрощают расчёт статистических характеристик таких как среднее значение и дисперсия. В принципе вычисление этих величин является несложной арифметической задачей. Однако, если значения анализируемой переменной являются «некрасивыми» числами, то для наборов большой длины даже расчёты с помощью калькулятора потребуют относительно много усилий. Применение для обработки информации электронных таблиц Microsoft Office Excel не решает все проблемы, т.к. возможны ошибки при вводе данных. Ниже мы докажем простые теоремы о линейных преобразованиях статистических переменных, которые во многих случаях позволяют упростить вычисления.

2.2 Изменение статистических характеристик при линейном преобразовании

Теорема 2.1. *Предположим, что числовая статистическая переменная Y получена из переменной X линейным преобразованием $Y = aX + b$. Тогда*

$$\mathrm{M}(Y) = a\,\mathrm{M}(X) + b, \tag{2.1}$$

$$\mathrm{D}(Y) = a^2\,\mathrm{D}(X), \tag{2.2}$$

$$\sigma(Y) = |a| \cdot \sigma(X). \tag{2.3}$$

Таким образом,

1. среднее значение $\mathrm{M}(Y)$ преобразованной переменной Y получается из среднего значения $\mathrm{M}(X)$ исходной переменной X с помощью той же линейной функции $y = ax + b$, которая преобразует значения переменной X в значения переменной Y;

2. дисперсия преобразованной переменной Y не зависит от параметра сдвига и пропорциональна квадрату масштабного параметра;

3. стандартное отклонение преобразованной переменной Y не зависит от параметра сдвига и пропорционально модулю масштабного параметра.

В частности, если масштабный множитель a равен 1 или -1, то $\mathrm{D}(Y) = \mathrm{D}(X)$, $\sigma(Y) = \sigma(X)$.

Доказательство. Следующая цепочка формул связывает величины $\mathrm{M}(X)$ и $\mathrm{M}(Y)$ между собой и доказывает соотношение (2.1):

$$\mathrm{M}(Y) = \frac{y_1 + \ldots + y_n}{n} = \frac{(ax_1 + b) + \ldots + (ax_n + b)}{n} = \frac{a(x_1 + \ldots + x_n) + nb}{n}$$

$$= a\frac{x_1 + \ldots + x_n}{n} + b = a\,\mathrm{M}(X) + b.$$

Для дисперсий мы аналогично имеем:

$$\begin{aligned}\mathrm{D}(Y) &= \frac{(y_1 - \mathrm{M}(Y))^2 + \ldots + (y_n - \mathrm{M}(Y))^2}{n} \\ &= \frac{(ax_1 + b - a\,\mathrm{M}(X) - b)^2 + \ldots + (ax_n + b - a\,\mathrm{M}(X) - b)^2}{n} \\ &= \frac{a^2(x_1 - \mathrm{M}(X))^2 + \ldots + a^2(x_n - \mathrm{M}(X))^2}{n} \\ &= a^2\,\frac{(x_1 - \mathrm{M}(X))^2 + \ldots + (x_n - \mathrm{M}(X))^2}{n} = a^2\,\mathrm{D}(X).\end{aligned}$$

Формулу (2.2) для дисперсий можно получить и с помощью ранее доказанной формулы (2.1) для средних значений, если воспользоваться формулой $\mathrm{D}(X) = \mathrm{M}\left(X^2\right) - \left(\mathrm{M}(X)\right)^2$:

$$\begin{aligned}\mathrm{D}(Y) &= \mathrm{M}\left(Y^2\right) - \left(\mathrm{M}(Y)\right)^2 = \mathrm{M}\left((aX+b)^2\right) - \left(\mathrm{M}(aX+b)\right)^2 \\ &= \mathrm{M}\left(a^2X^2 + 2abX + b^2\right) - \left(a\,\mathrm{M}(X) + b\right)^2 \\ &= a^2\,\mathrm{M}\left(X^2\right) + 2ab\,\mathrm{M}(X) + b^2 - \left(a^2\left(\mathrm{M}(X)\right)^2 + 2ab\,\mathrm{M}(X) + b^2\right) \\ &= a^2\left[\mathrm{M}\left(X^2\right) - \left(\mathrm{M}(X)\right)^2\right] = a^2\,\mathrm{D}(X).\end{aligned}$$

□

Замечание. Доказанную теорему можно сформулировать не на языке статистических переменных, а на языке наборов их значений: если набор $Y = \left[y_1, \ldots, y_n\right]$ получен из набора $X = \left[x_1, \ldots, x_n\right]$ линейным преобразованием $y_i = ax_i + b$, то верны равенства $\mathrm{M}(Y) = a\,\mathrm{M}(X) + b$, $\mathrm{D}(Y) = a^2\,\mathrm{D}(X)$, $\sigma(Y) = |a| \cdot \sigma(X)$.

Часто можно найти такое линейное преобразование исходного набора, что новый набор выглядит проще. Тогда найти среднее значение для нового набора легче, чем для исходного. Полученное в Теореме 2.1 общее соотношение между средними значениями исходного и нового наборов позволяет найти среднее значение исходного набора меньшими вычислительными усилиями.

Пример 2.2. Предположим, что основной набор X имеет вид: $\left[2.91,\ 3.07,\ 3.02,\ 3.05,\ 3.01,\ 3.03\right]$. Введём новый набор Y по формуле: $Y = 100X - 300$, так что $Y = \left[-9,\ 7,\ 2,\ 5,\ 1,\ 3\right]$. Среднее значение этого набора совсем несложно вычислить даже в уме:

$$\mathrm{M}(Y) = \frac{-9 + 7 + 2 + 5 + 1 + 3}{6} = \frac{9}{6} = 1.5\,.$$

Поскольку $X = 3 + 0.01Y$, аналогичное соотношение связывает средние значения: $\mathrm{M}(X) = 3 + 0.01\,\mathrm{M}(Y)$. Поэтому для среднего значения исходного набора мы имеем: $\mathrm{M}(X) = 3 + 0.01 \cdot 1.5 = 3.015$.

Равным образом легко вычислить и $\mathrm{M}\left(Y^2\right)$:

$$\mathrm{M}\left(Y^2\right)=\frac{9^2+7^2+2^2+5^2+1^2+3^2}{6}=\frac{169}{6}\ .$$

Теперь для дисперсии нового набора мы имеем:

$$\mathrm{D}(Y)=\mathrm{M}\left(Y^2\right)-\left(\mathrm{M}(Y)\right)^2=\frac{169}{6}-\frac{9}{4}=\frac{311}{12}\approx 25.92.$$

Соответственно, $\sigma(Y)\approx 5.091$. Применяя Теорему 2.1, мы имеем: $\sigma(X)=0.01\sigma(Y)\approx 0.0591$, $\mathrm{D}(X)=0.01^2\,\mathrm{D}(Y)\approx 0.002592$.

Теорема 2.2. *Предположим, что набор* $Y=\left[y_1,\ldots,y_n\right]$ *получен из набора* $X=\left[x_1,\ldots,x_n\right]$ *линейным преобразованием* $Y=aX+b$. *Тогда*

$$\mu(Y)=a\mu(X)+b, \tag{2.4}$$

$$R(Y)=\left|a\right|\cdot R(X). \tag{2.5}$$

Таким образом,

1. *медиана* $\mu(Y)$ *преобразованного набора Y получается из медианы* $\mu(X)$ *исходного набора X с помощью той же линейной функции* $y=ax+b$, *которая преобразует элементы набора X в элементы набора Y;*

2. *размах преобразованного набора Y не зависит от параметра сдвига и пропорционален модулю масштабного параметра.*

В частности, если масштабный множитель a равен 1 *или* -1, *то* $R(Y)=R(X)$

.

Доказательство. Если $x_i\le x_j$, то при $a>0$ аналогичное неравенство выполнено и для соответствующих элементов $y_i=ax_i+b$, $y_j=ax_j+b$ набора $Y=\left[y_1,\ldots,y_n\right]$: $y_i\le y_j$. Поэтому, если $\left(x_{(1)},\ldots,x_{(n)}\right)$ – упорядоченная по возрастанию версия набора X (т.е. вариационный ряд), то $\left(ax_{(1)}+b,\ldots,ax_{(n)}+b\right)$ – вариационный ряд для набора Y.

В частности, $y_{(1)}=ax_{(1)}+b$, $y_{(n)}=ax_{(n)}+b$, т.е. при линейном преобразовании с положительным масштабным параметром минимальный член набора X перейдёт в минимальный член набора Y, а максимальный член набора X перейдёт в максимальный член набора Y. Поэтому

$$R(Y)\equiv y_{(n)}-y_{(1)}=\left(ax_{(n)}+b\right)-\left(ax_{(1)}+b\right)=a\left(x_{(n)}-x_{(1)}\right)\equiv aR(X).$$

Далее, если $n=2k-1$ – нечётное число, то

$$\mu(Y)=y_{(k)}=ax_{(k)}+b=a\mu(X)+b.$$

Если $n=2k$ – чётное число, то

$$\mu(Y)=\frac{y_{(k)}+y_{(k+1)}}{2}=\frac{\left(ax_{(k)}+b\right)+\left(ax_{(k+1)}+b\right)}{2}=a\frac{x_{(k)}+x_{(k+1)}}{2}+b=a\mu(X)+b.$$

Таким образом, (2.4) доказано для $a>0$, как в случае $n=2k-1$, так и в случае $n=2k$.

Пусть теперь $a<0$. Тогда неравенство $x_i \le x_j$ для соответствующих элементов $y_i = ax_i + b$, $y_j = ax_j + b$ набора Y влечёт противоположное неравенство $y_i \ge y_j$. Поэтому, если $\left(x_{(1)},\ldots,x_{(n)}\right)$ – упорядоченная по возрастанию версия набора X (т.е. вариационный ряд), то вариационный ряд для набора Y имеет вид $\left(ax_{(n)}+b,\ldots,ax_{(1)}+b\right)$.

В частности, $y_{(1)} = ax_{(n)} + b$, $y_{(n)} = ax_{(1)} + b$, т.е. при линейном преобразовании с отрицательным масштабным параметром минимальный член набора X перейдёт в максимальный член набора Y, а максимальный член набора X перейдёт в минимальный член набора Y. Поэтому

$$R(Y) \equiv y_{(n)} - y_{(1)} = \left(ax_{(1)}+b\right) - \left(ax_{(n)}+b\right) = -a\left(x_{(n)} - x_{(1)}\right) \equiv |a| R(X).$$

Далее, нетрудно сообразить, что на k-м месте в последовательности $\left(ax_{(n)}+b,\ldots,ax_{(1)}+b\right)$ стоит число $ax_{(n-k+1)}+b$, т.е. $y_{(k)} = ax_{(n-k+1)} + b$. Поэтому, если $n = 2k-1$ – нечётное число, то

$$\mu(Y) = y_{(k)} = ax_{(n-k+1)} + b = ax_{(2k-1-k+1)} + b = ax_{(k)} + b = a\mu(X) + b.$$

Полученный результат очевиден из следующего простого наблюдения: если последовательность состоит из нечётного числа $n = 2k-1$ членов, то при изменении порядка членов последовательности на противоположный средний, k-й член, останется на месте.

Если же $n = 2k$ – чётное число, то

$$\mu(Y) = \frac{y_{(k)} + y_{(k+1)}}{2} = \frac{\left(ax_{(n-k+1)}+b\right) + \left(ax_{(n-k+1)}+b\right)}{2}$$
$$= \frac{\left(ax_{(2k-k+1)}+b\right) + \left(ax_{(2k-k+1)}+b\right)}{2} = a\frac{x_{(k+1)} + x_{(k)}}{2} + b = a\mu(X) + b.$$

Полученный результат очевиден из следующего простого наблюдения: если последовательность состоит из чётного числа $n = 2k$ членов, то при изменении порядка членов последовательности на противоположный средние члены, k-й и (k+1)-й, поменяются местами, но их сумма, а вместе с ней и среднее арифметическое, не изменится.

Замечание. Соотношения (2.4) и (2.5) полностью аналогичны соотношениям (2.1) и (2.3). Это вполне соответствует тому, что медиана, как и среднее, является мерой положения чисел набора, а размах, как и стандартное отклонение, является мерой разброса.

Пример 2.3. Вернёмся к Примеру 2.2, где мы рассматривали основной набор $X = \left[2.91,\ 3.07,\ 3.02,\ 3.05,\ 3.01,\ 3.03\right]$ и его линейное преобразование по формуле $Y = 100X - 300$, так что $Y = \left[-9,\ 7,\ 2,\ 5,\ 1,\ 3\right]$.

Вариационный ряд для исходного набора имеет вид:

2.91, 3.01, 3.02, 3.03, 3.05, 3.07.

Соответственно $\min\limits_{1\le i\le 6} x_i = x_{(1)} = 2.91$, $\max\limits_{1\le i\le 6} x_i = x_{(6)} = 3.07$ и потому размах $R(X) = x_{(6)} - x_{(1)} = 3.07 - 2.91 = 0.16$.

Вариационный ряд для преобразованного набора Y имеет вид:

$$-9,\ 1,\ 2,\ 3,\ 5,\ 7.$$

Соответственно $\min\limits_{1\le i\le 6} y_i = y_{(1)} = -9$, $\max\limits_{1\le i\le 6} y_i = y_{(6)} = 7$ и потому размах $R(Y) = y_{(6)} - y_{(1)} = 7 - (-9) = 16$.

Поскольку $X = 0.01Y + 3$, величины $R(X)$ и $R(Y)$, как и утверждает формула (2.5), связаны соотношением $R(X) = 0.01R(Y)$.

Для медиан мы имеем: $\mu(X) = \dfrac{x_{(3)} + x_{(4)}}{2} = \dfrac{3.02 + 3.03}{2} = 3.025$, $\mu(Y) = \dfrac{y_{(3)} + y_{(4)}}{2} = \dfrac{2 + 3}{2} = 2.5$. Поскольку $X = 0.01Y + 3$, величины $\mu(X)$ и $\mu(Y)$, как и утверждает формула (2.4), связаны соотношением: $\mu(X) = 0.01\mu(Y) + 3$.

Пример 2.4. В магазине продают семь видов сока в литровых упаковках по следующим ценам (в рублях):

$x_1 = 28.90, x_2 = 34.90, x_3 = 29.90, x_4 = 32.90, x_5 = 39.90, x_6 = 24.90, x_7 = 38.90.$

Определите среднюю цену литрового пакета сока, стандартное отклонение от этой цены, а также медиану и размах цен.

Решение. Все цены оканчиваются записью «.90» – это обычный приём розничной торговли, который позволяет создать впечатление, что товар дешевле, чем на самом деле. Например, для потребителя, в сущности, нет никакой разницы между 40 руб. и 39.90 руб. Однако психологически потребитель относит товар ценой 39.90 руб. в категорию 30 рублёвых товаров и может купить его, в то время как указание цены 40 рублей автоматически переводит товар в категорию 40 рублёвых товаров, что может быть неприемлемо для покупателя.

Имея в виду это соображение, введём округлённые цены $u_i = x_i + 0.10$, что даст набор $U = \left[29, 35, 30, 33, 40, 25, 39\right]$.

Значения этих округлённых цен колеблются вокруг числа 30. Поэтому введём ещё один набор $Y = \left[y_1, \ldots, y_7\right]$ с помощью преобразования $Y = U - 30$, так что $Y = \left[-1,\ 5,\ 0,\ 3,\ 10, -5,\ 9\right]$. Преобразование исходного набора X, которое приводит к набору Y, имеет вид: $Y = X - 29.90$.

Средние значения наборов Y и Y^2 легко подсчитать в уме: $\mathrm{M}(Y) = 3$, $\mathrm{M}\left(Y^2\right) = \dfrac{241}{7}$. Теперь для дисперсии набора Y мы имеем (все вычисления можно проделать без калькулятора): $\mathrm{D}(Y) = \dfrac{178}{7} \approx 25.43$. Соответственно, $\sigma(Y) = \sqrt{\mathrm{D}(Y)} \approx 5.04$.

Чтобы подсчитать медиану и размах, образуем вариационный ряд $\left(-5, -1, 0, 3, 5, 9, 10\right)$ вспомогательного набора Y. Размах этого набора равен $\mathrm{R}(Y) = 10 - (-5) = 15$. Далее, количество чисел в наборе Y (как и в исходном

наборе X) равно 7, т.е. является нечётным числом. Поэтому медиана набора Y – это четвёртый член соответствующего вариационного ряда, т.е. $\mu(Y) = 3$.

Теперь легко подсчитать статистические характеристики основного набора. Средние значения и медианы связаны тем же линейным соотношением $X = Y + 29.90$, которое связывает числа наборов X и Y: $\mathrm{M}(X) = \mathrm{M}(Y) + 29.90 = 32.90$, $\mu(X) = \mu(Y) + 29.90 = 32.90$, а поскольку масштабный коэффициент при преобразовании набора Y в набор X равен 1, их меры разброса совпадают: $\sigma(X) = \sigma(Y) \approx 5.04$, $\mathrm{R}(X) = \mathrm{R}(Y) = 15$.

2.3 Стандартизация статистических переменных

Пусть X – некоторая числовая статистическая переменная, $[x_1,\ldots,x_n]$ – набор её значений, $\mathrm{M}(X)$ – её среднее значение. Рассмотрим переменную $Y = [y_1,\ldots,y_n]$, полученную из переменной $X = [x_1,\ldots,x_n]$ применением линейного преобразования:

$$Y = X - \mathrm{M}(X) \Leftrightarrow y_i = x_i - \mathrm{M}(X),\ i = 1,\ldots,n. \qquad (2.6)$$

Поскольку среднее значение набора чисел в некотором смысле является его «центром», преобразование (2.6) называют *центрированием* исходного набора (или соответствующей переменной). Таким образом, среднее значение центрированной переменной равно 0.

Величины y_i показывают отклонения величин x_i от их среднего значения. Применяя формулу (2.1), мы получим: $\mathrm{M}(Y) = \mathrm{M}(X) - \mathrm{M}(X) = 0$. Таким образом, среднее значение (а вместе с ним и обычная сумма) отклонений величин x_i от их среднего значения обязательно равна нулю.

Пусть $X = [x_1,\ldots,x_n]$ – числовой набор с ненулевой дисперсией $\mathrm{D}(X)$. Рассмотрим набор $Y = [y_1,\ldots,y_n]$, полученный из набора $X = [x_1,\ldots,x_n]$ применением линейного преобразования

$$Y = \frac{X}{\sqrt{\mathrm{D}(X)}} \Leftrightarrow Y = \frac{X}{\sigma(X)}\ . \qquad (2.7)$$

Это преобразование называют *нормированием* исходного набора (или соответствующей переменной). Применяя формулу (2.2), мы получим: $\mathrm{D}(Y) = \mathrm{D}(X) \Big/ \left(\sqrt{\mathrm{D}(X)}\right)^2 = 1$. Таким образом, дисперсия (а вместе с ней и стандартное отклонение) нормированной переменной обязательно равно 1.

Пусть $X = [x_1,\ldots,x_n]$ – некоторый набор чисел, $\mathrm{M}(X)$ – его среднее значение, $\mathrm{D}(X)$ – дисперсия, $\sigma(X) = \sqrt{\mathrm{D}(X)}$ – среднее квадратичное отклонение. Предположим, что дисперсия отлична от нуля и рассмотрим набор $Y = [y_1,\ldots,y_n]$, полученный из набора $X = [x_1,\ldots,x_n]$ применением следующего линейного преобразования:

$$Y = \frac{X - \mathrm{M}(X)}{\sigma(X)}. \qquad (2.8)$$

Применяя Теорему 2.1, мы получим:

$$\mathrm{M}(Y) = \frac{\mathrm{M}(X) - \mathrm{M}(X)}{\sigma(X)} = 0, \quad \mathrm{D}(Y) = \frac{\mathrm{D}(X)}{\left(\sigma(X)\right)^2} = 1.$$

Статистическая переменная (набор) с нулевым средним и единичной дисперсией называется стандартной. Соответственно, преобразование (2.8) называют *стандартизацией* исходной переменной (набора). Оно играет важную роль в теории вероятностей и статистике.

В статистике стандартизованные наборы могут использоваться для того, чтобы сравнивать разные наборы данных, которые были получены в результате измерения характеристик одних и тех же объектов с помощью разных процедур и разных единиц измерения. Поэтому прямое сравнение чисел из таких наборов обычно не имеет смысла. Стандартизованный, т.е. центрированный и нормированный, набор $Z = \left(\frac{x_1 - \mathrm{M}(X)}{\sigma(X)}, \ldots, \frac{x_n - \mathrm{M}(X)}{\sigma(X)} \right)$ измеряет величины отклонений чисел x_i от их среднего $M(X)$ не в абсолютных значениях, а по отношению к стандартному отклонению $\sigma(X)$ (которое тем самым выбирается в качестве новой единицы измерения величин x_i). Хотя далеко не всегда стандартизация набора является наилучшим подходом к задаче сравнения чисел из разных наборов, она очень проста как с методической, так и с вычислительной точек зрения.

Чтобы проиллюстрировать эти общие рассуждения, проанализируем гипотетические результаты ЕГЭ по математике для шести школьников, приведённые во второй строке Таблицы 2.1 (максимально возможное число баллов равно 100):

Таблица 2.1

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Исходная оценка, x_i	71	72	74	75	93	95
Центрированная оценка	–9	–8	–6	–5	13	15
Стандартизованная оценка, z_i	–0.9	–0.8	–0.6	–0.5	1.3	1.5

`

Сумма всех шести оценок равна 480, так что средняя арифметическая оценка равна $\mathrm{M}(X) = \frac{480}{6} = 80$. Отклонения оценок от этого среднего (центрированные оценки за экзамен) приведены в третьей строке Таблицы 2.1. Центрированные оценки позволяют легко определить, хуже или лучше результат школьника, чем средний результат: если центрированная оценка отрицательна, то результат хуже среднего, а если положительна – то лучше. Например, центрированная оценка Андрея равна 13. Поэтому его результат выше среднего (по рассматриваемой группе из 6 школьников).

Сумма квадратов отклонений оценок от среднего равна 600, так что дисперсия исходного набора оценок (которая по определению равна среднему

значению квадратов отклонений оценок от среднего) есть $\mathrm{D}(X)=\frac{600}{6}=100$. Соответственно, стандартное отклонение для исходного набора оценок равно $\sigma(X)=\sqrt{100}=10$. Теперь мы может подсчитать стандартизованные оценки $z_i=\frac{x_i-\mathrm{M}(X)}{\sigma(X)}=\frac{x_i-80}{10}$. Они приведены в последней строке Таблицы 2.1 и могут рассматриваться как «места», которые занимают школьники *данной* группы по итогам *данного* экзамена. Иначе говоря, стандартизованные оценки позволяют количественно описать разрыв между школьниками рассматриваемой группы по итогам конкретного экзамена с учётом степени разброса всех оценок.

Предположим теперь, что эти ребята решили поступать в университет, который проводит дополнительный экзамен по математике, и получили оценки, которые приведены во второй строке Таблицы 2.2 (максимально возможное число баллов равно 100):

Таблица 2.2

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Исходная оценка, x_i'	60	57	50	50	51	50
Центрированная оценка	7	4	–3	–3	–2	–3
Стандартизованная оценка	1.75	1	–0.75	–0.75	–0.5	–0.75

Если зачисление проводится по результатам ЕГЭ и дополнительного вступительного экзамена, то обычно складывают оценки по двум экзаменам и затем отбирают столько лучших абитуриентов, сколько имеется мест. В нашем случае суммарные оценки приведены в Таблице 2.3.

Таблица 2.3

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
Сумма оценок	131	129	124	125	144	145

Если, например, эти шесть абитуриентов претендуют на одно место, то зачислена в университет будет Оля, которая опередила ближайшего конкурента, Андрея, на 1 балл, а следующего, Петю, на 14 баллов.

Проанализируем теперь эту ситуацию с помощью стандартизованных оценок. Нетрудно подсчитать, что для дополнительного экзамена сумма всех шести оценок равна 318, так что средняя арифметическая оценка равна $\mathrm{M}(X')=\frac{318}{6}=53$. Отклонения оценок от этого среднего (центрированные оценки за экзамен) приведены в третьей строке Таблицы 2.2.

Сумма квадратов отклонений оценок от среднего равна 96, так что $\mathrm{D}(X')=\frac{96}{6}=16$. Соответственно, стандартное отклонение для оценок за ДВИ равно $\sigma(X')=\sqrt{16}=4$. Теперь мы может подсчитать стандартизованные оценки $z_i'=\frac{x_i'-\mathrm{M}(X')}{\sigma(X')}=\frac{x_i'-53}{4}$; они приведены в последней строке Таблицы 2.2.

Стандартизованная оценка по ЕГЭ Маши и стандартизованная оценка по ДВИ Андрея совпадают (обе равны –0,5), хотя исходные оценки совершенно разные (75 у Маши, 51 у Андрея) и отличаются от средней за соответствующий экзамен по разному: у Маши оценка хуже средней на 5 баллов, а у Андрея оценка хуже средней на 2 балла. Но с точки зрения сложности экзамена и критериев оценок (для данной группы школьников) эти отклонения равноценны (оба равны половине соответствующего стандартного отклонения). Поэтому складывать первичные баллы не очень разумно. Переходя к стандартизованным оценкам, мы вводим более объективные показатели уровня знаний школьников на *конкретном экзамене по отношению к уровню знаний других школьников данной группы*.

Сумма стандартизованных оценок $z_i + z_i'$ приведена в Таблице 2.4.

Таблица 2.4

Имя школьника	Петя	Коля	Таня	Маша	Андрей	Оля
$z_i + z_i'$	0.85	0.2	–1.35	–1.25	0.8	0.75

По данным Таблицы 2.4 лучшая суммарная оценка у Пети и поэтому именно он, а не Оля (которая оказалась на третьем месте), должен быть зачислен в университет.

Решение о том, какой способ определения суммарной оценки следует применять при решении вопроса о зачислении в университет, лежит вне математики. В реальности приёмные комиссии просто суммируют оценки за разные экзамены. Однако в подобных ситуациях, когда характер экзаменов и принципы выставления оценок сильно отличаются, статистика рекомендует использовать стандартизованные оценки или другие аналогичные процедуры.

Чтобы ещё раз проиллюстрировать применение стандартизованных статистических переменных, разберём две дополнительные задачи.

Первая задача предлагалась британским экзаменационным центром Edexcel в июне 2008 г. на выпускном школьном экзамене по курсу GCSE Statistics (задача №6; для удобства восприятия перевод оригинального английского условия задачи немного отредактирован).

Задача 1. В конце четверти в классе были проведены тесты по статистике и математике. Максимальное число баллов за каждый тест равно 100. Анализируя результаты этих тестов учитель установил, что

- по статистике средний балл равен 52, а стандартное отклонение равно 15,
- по математике средний балл равен 45, а стандартное отклонение равно 12.

(1) Прокомментируйте эти результаты.

Джон заработал 55 баллов по статистике и 48 баллов по математике.

(2) Подсчитайте стандартизованные оценки Джона по этим тестам.

(3) В каком предмете его успехи выше? Аргументируйте ваш ответ.

Решение. (1) Если мы посмотрим на результаты всего класса, то можно отметить, что средняя оценка по статистике выше средней оценки по математике. Это означает, что математика – более трудный предмет или тест по математике состоял из более сложных задач. Возможно, впрочем, что

школьники меньше внимания уделяли математике или проходили сложные темы, которые трудно было хорошо понять. Различие в разбросе оценок также указывает на то, что степень усвоения материала школьниками и/или уровень сложности тестов по этим двум предметам различны. Поэтому сопоставление абсолютных значений оценок по эти предметам нельзя считать разумным. Простейший способ учесть отмеченные выше обстоятельства связан с использованием стандартизованных оценок.

(2) Стандартизованная оценка Джона по статистике равна

$$\frac{\text{первичный балл по статистике} - \text{средний балл по классу}}{\text{стандартное отклонение баллов учеников от среднего}} = \frac{55-52}{15} = 0.2.$$

Стандартизованная оценка Джона по математике равна

$$\frac{\text{первичный балл по математике} - \text{средний балл по классу}}{\text{стандартное отклонение баллов учеников от среднего}} = \frac{48-45}{12} = 0.25.$$

(3) Поскольку стандартизованная оценка Джона по математике больше, чем его стандартизованная оценка по статистике, мы должны признать, что по математике успехи Джона выше.

На первый взгляд этот вывод противоречит данным тестов: ведь по статистике Джон заработал на 7 баллов больше, чем по математике. Однако, как мы отмечали, тест по математике оказался для всего класса сложнее теста по статистике. Поэтому оценивая успехи Джона нужно принимать в расчёт общую успеваемость в классе. Отклонение оценки Джона по математике от средней по классу оценки по этому предмету равно $48-45=3$, т.е. такое же, как и отклонение оценки Джона по статистике от средней по классу оценки по статистике ($55-52=3$). Но для всего набора оценок по статистике стандартное отклонение больше, чем стандартное отклонение для всего набора оценок по математике. Поэтому более высокий балл Джона по статистике может быть связан и с бóльшими колебаниями оценок по статистике. Чтобы нивелировать это различие, нужно измерять отклонения оценки по предмету от средней оценки не абсолютным числом баллов, а по отношению к типичному отклонению, одной из мер которого является стандартное отклонение.

Следующая задача взята из британского школьного учебника по статистике A.Ballard, S.Gill, et al. *GCSE Statistics. Complete Revision and Practice*. Coordination Group Publications Ltd, 2010 (стр. 112; для удобства восприятия перевод оригинального английского условия задачи немного отредактирован).

Задача 2. В соревнованиях по фигурному катанию на льду принимали участие 20 спортсменов. Каждый из них должен был исполнить 2 танца – обязательный и произвольный. За каждый танец участник может получить максимум 50 баллов. По результатам соревнований оказалось, что для обязательного танца средняя оценка по группе равна 37.25, а стандартное отклонение оценки от средней равно 5.07. Для произвольного танца средняя оценка по группе равна 41.7, а стандартное отклонение оценки от средней равно 3.2.

(1) Участник A получил за исполнение обязательного танца 40 баллов, а за исполнение произвольного танца – 44 балла. Найдите соответствующие стандартизованные оценки и их сумму.

(2) Участник B получил за исполнение обязательного танца 43 балла, а за исполнение произвольного танца – 41 балл. Найдите соответствующие стандартизованные оценки и их сумму.

(3) Кто из этих двух участников показал лучший общий результат? Аргументируйте ваш ответ.

(4) Стандартизованная оценка участника C за исполнение обязательного танца равна 0.15. Сколько первичных баллов он получил за этот танец?

Решение.

(1) Стандартизованная оценка участника A за исполнение обязательного танца равна

$$\frac{\text{первичный балл участика A} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{40 - 37.25}{5.07} \approx 0.542.$$

Стандартизованная оценка участника A за исполнение произвольного танца равна

$$\frac{\text{первичный балл участика A} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{44 - 41.7}{3.2} \approx 0.719.$$

(2) Стандартизованная оценка участника B за исполнение обязательного танца равна

$$\frac{\text{первичный балл участика B} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{43 - 37.25}{5.07} \approx 1.134.$$

Стандартизованная оценка участника B за исполнение произвольного танца равна

$$\frac{\text{первичный балл участика B} - \text{средний балл}}{\text{стандартное отклонение}} = \frac{41 - 41.7}{3.2} \approx -0.219.$$

(3) Сумма стандартизованных оценок участника A приближённо равна 1.26. Для участника B эта сумма приближённо равна 0.92. Поэтому следует признать, что участник A опередил участника B (несмотря на то, что простая сумма первичных баллов у этих участников одна и та же; она равна 84).

(4) Из формулы $z_i = \dfrac{x_i - \mathrm{M}(X)}{\sigma(X)}$ мы имеем: $x_i = \mathrm{M}(X) + \sigma(X) \cdot z_i$. Поэтому первичный балл участника C за исполнение обязательного танца равен $37.25 + 0.15 \times 5.07 \approx 38.$

2.4 О нелинейных преобразованиях

Пусть X – статистическая переменная, все значения $\left[x_1, \ldots, x_n\right]$ которой отличны от 0. Если для каждого объекта i генеральной совокупности, объекты которой характеризует переменная X, рассмотреть число $y_i = 1/x_i$, то мы получим новую переменную Y, значения которой связаны со значениями переменной X формулой: $Y(i) = 1/X(i)$, $i = 1, \ldots, n$. Эту переменную мы будем обозначать $1/X$. Как обычно, вместо числовых переменных можно говорить о числовых наборах.

Замена переменной X на переменную $1/X$ является простейшим примером нелинейного преобразования. Свойства нелинейных преобразований кардинально отличаются от свойств линейных. В частности, если исключить тривиальный случай, когда переменная X принимает одно значение для всех объектов генеральной совокупности, и предположить, что все $x_i > 0$, то

$$\mathrm{M}\left(\frac{1}{X}\right) > \frac{1}{\mathrm{M}(X)} \ . \tag{2.9}$$

Поскольку $\mathrm{M}(1/X)$ является величиной, обратной среднему гармоническому чисел x_i, неравенство (2.9) фактически является классическим неравенством между средним арифметическим и средним гармоническим.

Различие между левой и правой частями неравенства (2.9) может быть поразительным. Рассмотрим, например, величину X с распределением значений, приведённым в Таблице 2.5.

Таблица 2.5

значение, y	0.01	0.1	1	10	100
абсолютная частота, $t(y)$	70	20	7	2	1
относительная частота, $f(y) = t(y)/n$	0.70	0.20	0.07	0.02	0.01

Иначе говоря, мы рассматриваем набор из 100 чисел, в котором число 0.01 повторяется 70 раз, число 0.1 – 20 раз, число 1 – 7 раз, число 10 – 2 раза, число 100 – 1 раз. Тогда

$$M(X) = 0.01 \cdot 0.7 + 0.1 \cdot 0.2 + 1 \cdot 0.07 + 10 \cdot 0.02 + 100 \cdot 0.01 = 1.297,$$

так что $1/\mathrm{M}(X) \approx 0.77$.

Величина $1/X$ имеет распределение значений, указанное в Таблице 2.6.

Таблица 2.6

значение, y	100	10	1	0.1	0.01
абсолютная частота, $t(y)$	70	20	7	2	1
относительная частота, $f(y) = t(y)/n$	0.70	0.20	0.07	0.02	0.01

Тогда

$$\mathrm{M}\left(\frac{1}{X}\right) = 100 \cdot 0.7 + 10 \cdot 0.2 + 1 \cdot 0.07 + 0.1 \cdot 0.02 + 0.01 \cdot 0.01 = 72.0721.$$

Таким образом, $\mathrm{M}\left(\frac{1}{X}\right)$ почти в 100 раз больше, чем $\frac{1}{\mathrm{M}(X)}$.

3 Суммирование статистических переменных

3.1 Операции над статистическими переменными

В предыдущем разделе мы говорили об определённых операциях над статистическими переменными – умножении на число и прибавлении числа. Там же были отмечены два важных результата:

$$\mathrm{M}(aX+b) = a\,\mathrm{M}(X) + b, \tag{3.1}$$

$$\mathrm{D}(aX+b) = a^2\,\mathrm{D}(X). \tag{3.2}$$

В результате применения линейного преобразования $Y = aX + b$ мы из *одной* статистической переменной, X, получаем тоже *одну* статистическую переменную, $Y = aX + b$. В математике такие операции называют *унарными*.

Обычные операции над числами (сложение, вычитание, умножение, деление) строят новое число (результат операции) по *двум* данным числам (в случае деления делитель должен быть отличен от нуля). В математике такие операции называют *бинарными*.

Соответственно, если мы говорим о сумме X+Y, разности $X - Y$, произведении $X \cdot Y$ двух статистических переменных, X и Y, то это должна быть однозначно определённая статистическая переменная. Если принять эту точку зрения, то мы с необходимостью приходим к важному выводу: складывать, вычитать и т.д. можно только числовые статистические переменные, характеризующие объекты одной и той же генеральной совокупности, т.е., в сущности, только компоненты двумерной переменной (X,Y).

Итак, пусть (X,Y) – двумерная числовая переменная. Это означает, что для каждого объекта i некоторой генеральной совокупности известны значения этих переменных, x_i и y_i. Тогда переменная $S = X + Y$ для объекта i принимает значение $s_i = x_i + y_i$, переменная $R = X - Y$ для объекта i принимает значение $r_i = x_i - y_i$, переменная $P = X \cdot Y$ для объекта i принимает значение $p_i = x_i \cdot y_i$, переменная $K = X^2$ для объекта i принимает значение $k_i = x_i^2$, переменная $F = \dfrac{X}{Y}$ для объекта i принимает значение $f_i = \dfrac{x_i}{y_i}$ (таким образом, $F = \dfrac{X}{Y}$ определена тогда и только тогда, когда *все* значения переменной Y отличны от нуля). Итак, мы естественным образом можем определить основные арифметические операции над компонентами двумерной статистической переменной.

Пусть, например, генеральная совокупность состоит из трёх школьников с именами Петя, Саша, Миша, переменная X характеризует число решённых ими

задач с кратким ответом из первой и второй частей ЕГЭ 2017 г. (всего 12 заданий): скажем, X(Петя)=8, X(Саша)=11, X(Миша)=12, а переменная Y характеризует число решённых ими задач с развёрнутым ответом из второй части ЕГЭ (всего 7 заданий): скажем, Y(Петя)=1, Y(Саша)=0, Y(Миша)=2. Тогда сумма $S = X + Y$ имеет простой смысл – это общее число задач, решённых школьником: S(Петя)=9, S(Саша)=11, S(Миша)=14.

Предположим теперь, что переменные X и Y измеряют некоторые свойства двух *разных* генеральных совокупностей. Пусть, например, первая совокупность состоит из трёх школьников с именами Петя, Саша, Миша, а переменная X характеризует число решённых ими задач с кратким ответом из первой и второй частей ЕГЭ 2017 г.: X(Петя)=8, X(Саша)=11, X(Миша)=12. Вторая совокупность состоит из трёх школьниц с именами Катя, Даша, Маша, а переменная Y характеризует число решённых ими задач с развёрнутым ответом из второй части ЕГЭ: Y(Катя)=1, Y(Даша)=0, Y(Маша)=2. Ясно, что в этой ситуации никакого содержательного смысла приписать «сумме» $X+Y$ нельзя. Если бы мы заменили переменные соответствующими наборами их значений, т.е. рассматривали X как неупорядоченный набор [8,11,12], а Y как неупорядоченный набор [1,0,2], то можно было бы объявить суммой этих наборов набор [8+1,11+0,12+2]=[9,11,14]. Однако, поскольку статистические наборы не упорядочены (так что [1,0,2]=[2,1,0]), с равным основанием можно было бы объявить суммой этих наборов и набор [8+2,11+1,12+0]=[10,12,12]. Решить эту проблему с неоднозначностью очень легко – нужно просто признать, что *складывать обычные (одномерные) наборы, относящиеся к разным генеральным совокупностям, нельзя*.

Однако, для наборов, относящихся к непересекающимся генеральным совокупностям, можно естественно определить другую бинарную операцию, не менее важную, чем сумма переменных – *объединение* наборов (на другом языке – *смесь распределений*). Об этой операции мы будем говорит в следующем разделе.

Для действий $+,-,\times,:$ над статистическими переменными, заданными на одной и той же генеральной совокупности, т.е. являющимися компонентами некоторой двумерной, или, в зависимости от ситуации, трёхмерной и т.д, переменной, выполнены обычные свойства аналогичных действий над числами: $X + Y = Y + X$, $X \cdot Y = Y \cdot X$, $X \cdot (Y + Z) = X \cdot Y + X \cdot Z$ и т.д. При этом роль нуля играет переменная O, которая для всех объектов равна 0 (соответствующий набор состоит из одних нулей), роль единицы играет переменная E, которая для всех объектов равна 1 (соответствующий набор состоит из одних единиц), так что $X + O = X$, $X \cdot E = X$.

Проверим, например, свойство $X \cdot (Y + Z) = X \cdot Y + X \cdot Z$. Если генеральная совокупность состоит из n объектов, то набор $Y + Z$ состоит из чисел $y_i + z_i$, $i = 1,\ldots,n$, а набор $X \cdot (Y + Z)$ – из чисел $x_i\left(y_i + z_i\right)$, $i = 1,\ldots,n$. С другой стороны, наборы $X \cdot Y$ и $X \cdot Z$ состоят из чисел $x_i y_i$ и $x_i z_i$, $i = 1,\ldots,n$, соответственно, а набор $X \cdot Y + X \cdot Z$ – из чисел $x_i y_i + x_i z_i$, $i = 1,\ldots,n$. Поскольку $x_i\left(y_i + z_i\right) = x_i y_i + x_i z_i$, можно утверждать, что $X \cdot (Y + Z) = X \cdot Y + X \cdot Z$.

Из этих общих свойств действий $+,-,\times,:$ над статистическими переменными вытекают другие формулы, к которым мы привыкли в алгебре чисел, например, формула $(X + Y)^2 = X^2 + 2XY + Y^2$. Доказательство этой

формулы для чисел использует только общие свойства действий $+,\times$; природа объектов, над которыми эти действия производятся, не играет никакой роли. Поэтому она верна и для статистических переменных.

3.2 Среднее значение суммы статистических переменных

Пусть (X,Y) – двумерная числовая переменная, которая для каждого объекта i некоторой генеральной совокупности из n объектов численно характеризует два его свойства. Как обычно, значения этих переменных для объекта i обозначим через x_i и y_i соответственно. Тогда переменная $S = X + Y$ для объекта i принимает значение $s_i = x_i + y_i$. Поэтому непосредственно по определению среднего значения мы имеем:

$$\mathrm{M}(S) = \frac{s_1 + \ldots + s_n}{n} = \frac{(x_1 + y_1) + \ldots + (x_n + y_n)}{n}.$$

Поскольку от перестановки слагаемых сумма не меняется,

$$\mathrm{M}(S) = \frac{(x_1 + \ldots + x_n) + (y_1 + \ldots + y_n)}{n} = \frac{x_1 + \ldots + x_n}{n} + \frac{y_1 + \ldots + y_n}{n}.$$

Применяя ещё раз определение среднего значения, мы получим: $\mathrm{M}(S) = \mathrm{M}(X) + \mathrm{M}(Y)$. Итак, нами доказана

Теорема 3.1. *Среднее значение суммы переменных равно сумме средних значений слагаемых:*

$$\mathrm{M}(X + Y) = \mathrm{M}(X) + \mathrm{M}(Y). \qquad (3.3)$$

С помощью (3.1) можно получить и более общее соотношение:

$$\mathrm{M}(aX + bY) = a\,\mathrm{M}(X) + b\,\mathrm{M}(Y). \qquad (3.4)$$

Для доказательства достаточно сначала применить (3.3) к переменным $X' = aX$ и $Y' = bY$, а затем воспользоваться соотношением (3.1):

$$\mathrm{M}(aX + bY) = \mathrm{M}(aX) + \mathrm{M}(bY) = a\,\mathrm{M}(X) + b\,\mathrm{M}(Y).$$

Соотношение, аналогичное (3.3), справедливо и для разности $R = X - Y$:

$$\mathrm{M}(X - Y) = \mathrm{M}(X) - \mathrm{M}(Y). \qquad (3.5)$$

Доказательство можно провести по аналогии с выводом (3.3):

$$\mathrm{M}(R) = \frac{r_1 + \ldots + r_n}{n} = \frac{(x_1 - y_1) + \ldots + (x_n - y_n)}{n} = \frac{(x_1 + \ldots + x_n) - (y_1 + \ldots + y_n)}{n}$$

$$= \frac{x_1 + \ldots + x_n}{n} - \frac{y_1 + \ldots + y_n}{n} = \mathrm{M}(X) - \mathrm{M}(Y).$$

С равным успехом для доказательства (3.5) можно использовать (3.4):

$$\mathrm{M}(X - Y) = \mathrm{M}(X + (-1)\cdot Y) = \mathrm{M}(X) + (-1)\cdot\mathrm{M}(Y) = \mathrm{M}(X) - \mathrm{M}(Y).$$

Замечание. По аналогии с (3.3) и (3.5) можно было бы предположить, что $\mathrm{M}(X\cdot Y)=\mathrm{M}(X)\cdot\mathrm{M}(Y)$, но на самом деле это соотношение неверно. В качестве контрпримера рассмотрим двумерный набор (X,Y)=[(2.0),(0,2)] из $n=2$ пар (так что генеральная совокупность состоит из двух объектов). Тогда:

- $X=[2,0]$, так что $\mathrm{M}(X)=1$,
- $Y=[0,2]$, так что $\mathrm{M}(Y)=1$,
- $X\cdot Y=[2\cdot 0,0\cdot 2]=[0,0]$, так что $\mathrm{M}(X\cdot Y)=0$.

Отметим также, что, т.к. наборы неупорядочены, Y можно записать и как [2,0], т.е статистически наборы X и Y неотличимы, хотя верно точное равенство $X=2-Y$.

3.3 Дисперсия суммы статистических переменных

Пусть опять (X,Y) – двумерная числовая переменная, которая для каждого объекта i некоторой генеральной совокупности из n объектов численно характеризует два его свойства, x_i и y_i – значения этих переменных для объекта i, $s_i=x_i+y_i$ – значение переменной $S=X+Y$ для объекта i.

Для этой же генеральной совокупности можно ввести центрированные переменные $X-\mathrm{M}(X)$, $Y-\mathrm{M}(Y)$ и их произведение $\big(X-\mathrm{M}(X)\big)\cdot\big(Y-\mathrm{M}(Y)\big)$ – для объекта i они принимают значения $x_i-\mathrm{M}(X)$, $y_i-\mathrm{M}(Y)$, $\big(x_i-\mathrm{M}(X)\big)\cdot\big(y_i-\mathrm{M}(Y)\big)$ соответственно.

Непосредственно по определению дисперсии мы имеем:

$$\mathrm{D}(X+Y)\equiv\mathrm{D}(S)=\mathrm{M}\Big(\big[S-\mathrm{M}(S)\big]^2\Big)\equiv\mathrm{M}\Big(\big[X+Y-\mathrm{M}(X+Y)\big]^2\Big).$$

Используя формулу (3.3), мы получим:

$$\mathrm{D}(X+Y)=\mathrm{M}\Big(\big[X+Y-\mathrm{M}(X)-\mathrm{M}(Y)\big]^2\Big)=\mathrm{M}\Big(\big[\big(X-\mathrm{M}(X)\big)+\big(Y-\mathrm{M}(Y)\big)\big]^2\Big).$$

Возведём в квадрат выражение в фигурных скобках:

$$\mathrm{D}(X+Y)=\mathrm{M}\Big(\big[X-\mathrm{M}(X)\big]^2+\big[Y-\mathrm{M}(Y)\big]^2+2\big[X-\mathrm{M}(X)\big]\cdot\big[Y-\mathrm{M}(Y)\big]\Big)$$

и применим формулу (3.3) и (2.1):

$\mathrm{D}(X+Y)=\mathrm{M}\Big(\big[X-\mathrm{M}(X)\big]^2\Big)+\mathrm{M}\Big(\big[Y-\mathrm{M}(Y)\big]^2\Big)+2\,\mathrm{M}\Big(\big[X-\mathrm{M}(X)\big]\cdot\big[Y-\mathrm{M}(Y)\big]\Big)$. Первые два слагаемых в правой части (непоследственно по определению) равны $\mathrm{D}(X)$ и $\mathrm{D}(Y)$ соответственно. Появившееся дополнительное выражение $\mathrm{M}\Big(\big[X-\mathrm{M}(X)\big]\cdot\big[Y-\mathrm{M}(Y)\big]\Big)$ называется *ковариацией* переменных X и Y и обозначается $\mathrm{Cov}(X;Y)$:

$$\mathrm{Cov}(X;Y)\equiv\mathrm{M}\Big(\big[X-\mathrm{M}(X)\big]\cdot\big[Y-\mathrm{M}(Y)\big]\Big). \qquad (3.6)$$

Ковариация тесно связана с понятием статистической независимости.

Резюмируя проведённые выкладки, мы получим следующий важный результат для дисперсии суммы.

Теорема 3.2. *Дисперсия суммы переменных равна сумме дисперсий слагаемых, плюс удвоенная ковариация переменных:*

$$\mathrm{D}(X+Y) = \mathrm{D}(X) + \mathrm{D}(Y) + 2\cdot\mathrm{Cov}(X;Y). \tag{3.7}$$

Если $\mathrm{Cov}(X;Y) = 0$, то переменные X и Y называются *некоррелированными*. Для некоррелированных переменных дисперсия суммы равна сумме дисперсий.

Определению ковариации (3.6) можно придать другой вид. Раскрывая скобки в выражении в фигурных скобках в правой части и применяя (3.4), мы получим:

$$\begin{aligned}\mathrm{Cov}(X;Y) &= \mathrm{M}\big(X\cdot Y - \mathrm{M}(X)\cdot Y - \mathrm{M}(Y)\cdot X + \mathrm{M}(X)\cdot\mathrm{M}(Y)\big)\\ &= \mathrm{M}\big(X\cdot Y\big) - \mathrm{M}(X)\cdot\mathrm{M}(Y) - \mathrm{M}(Y)\cdot\mathrm{M}(X) + \mathrm{M}(X)\cdot\mathrm{M}(Y)\\ &= \mathrm{M}\big(X\cdot Y\big) - \mathrm{M}(X)\cdot\mathrm{M}(Y).\end{aligned} \tag{3.8}$$

Замечание в конце предыдущего раздела показывает, что, вообще говоря, $\mathrm{Cov}(X;Y) \neq 0$. Для дальнейшего отметим два простых свойства ковариации.

Прежде всего, непосредственно из определения этого понятия вытекает, что:

$$\mathrm{Cov}(X;Y) = \mathrm{Cov}(Y;X), \tag{3.9}$$

т.е. ковариация не зависит от порядка, в котором записаны переменные.

Кроме того,

$$\begin{aligned}\mathrm{Cov}(X_1+X_2,Y) &= \mathrm{M}\big((X_1+X_2)\cdot Y\big) - \mathrm{M}(X_1+X_2)\cdot\mathrm{M}(Y)\\ &= \mathrm{M}\big(X_1Y + X_2Y\big) - \big(\mathrm{M}(X_1) + \mathrm{M}(X_2)\big)\cdot\mathrm{M}(Y)\\ &= \mathrm{M}\big(X_1Y\big) + \mathrm{M}\big(X_2Y\big) - \mathrm{M}(X_1)\cdot\mathrm{M}(Y) - \mathrm{M}(X_2)\cdot\mathrm{M}(Y)\\ &= \mathrm{M}\big(X_1Y\big) - \mathrm{M}(X_1)\cdot\mathrm{M}(Y) + \mathrm{M}\big(X_2Y\big) - \mathrm{M}(X_2)\cdot\mathrm{M}(Y)\\ &= \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y),\end{aligned} \tag{3.10}$$

$$\begin{aligned}\mathrm{Cov}(aX,Y) &= \mathrm{M}(aX\cdot Y) - \mathrm{M}(aX)\cdot\mathrm{M}(Y) = a\,\mathrm{M}(XY) - a\,\mathrm{M}(X)\cdot\mathrm{M}(Y)\\ &= a\cdot\big(\mathrm{M}(XY) - \mathrm{M}(X)\cdot\mathrm{M}(Y)\big) = a\cdot\mathrm{Cov}(X,Y).\end{aligned} \tag{3.11}$$

Если дисперсии переменных X и Y отличны от нуля, т.е. эти переменные принимают хотя бы два значения (это наиболее интересный для практики случай), то можно ввести величину

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{D}(X)\cdot\mathrm{D}(Y)}} \equiv \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}. \tag{3.12}$$

Её называют *коэффициентом корреляции* переменных (или между переменными) X и Y. В терминах коэффицента корреляции соотношение (3.7) примет вид:

$$\mathrm{D}(X+Y) = \mathrm{D}(X) + \mathrm{D}(Y) + 2\rho(X;Y)\sqrt{\mathrm{D}(X)\cdot\mathrm{D}(Y)}. \tag{3.13}$$

3.4 Неравенство для ковариации

Теорема 3.3. *Для любой двумерной переменной верно неравенство*

$$\left|\mathrm{Cov}(X,Y)\right| \le \sqrt{\mathrm{D}(X)\cdot\mathrm{D}(Y)} \Leftrightarrow \left|\mathrm{Cov}(X,Y)\right| \le \sigma(X)\cdot\sigma(Y), \tag{3.14}$$

или, что то же самое, двойное неравенство

$$-\sigma(X)\cdot\sigma(Y) \le \mathrm{Cov}(X;Y) \le \sigma(X)\cdot\sigma(Y) \tag{3.15}$$

Соответственно, если дисперсии $\mathrm{D}(X)$ *и* $\mathrm{D}(Y)$ *отличны от нуля, то коэффициент корреляции удовлетворяет двойному неравенству*

$$-1 \le \rho(X,Y) \le 1. \tag{3.16}$$

Доказательство. Рассмотрим переменную $Z = tX + Y$ (t – некоторый вспомогательный числовой параметр) и подсчитаем её дисперсию. Используя доказанные выше свойства дисперсии и ковариации, мы имеем:

$$\mathrm{D}(tX+Y) = \mathrm{D}(tX) + \mathrm{D}(Y) + 2\cdot\mathrm{Cov}(tX;Y) = t^2\,\mathrm{D}(X) + 2t\,\mathrm{Cov}(X;Y) + \mathrm{D}(Y).$$

Если $\mathrm{D}(X) \ne 0$, то выражение в правой части является квадратным трёхчленом относительно t. Этот трёхчлен неотрицателен при всех t (он равен дисперсии $\mathrm{D}(tX+Y)$, которая неотрицательна как всякая дисперсия). Значит, его дискриминант отрицателен или равен 0:

$$4\left[\mathrm{Cov}(X,Y)\right]^2 - 4\,\mathrm{D}(X)\cdot\mathrm{D}(Y) \le 0$$
$$\Updownarrow$$
$$\left[\mathrm{Cov}(X;Y)\right]^2 \le \mathrm{D}(X)\cdot\mathrm{D}(Y)$$
$$\Updownarrow$$
$$\left|\mathrm{Cov}(X;Y)\right| \le \sqrt{\mathrm{D}(X)\cdot\mathrm{D}(Y)}.$$

Если $\mathrm{D}(X) = 0$, то переменная X для всех объектов анализируемой совокупности принимает одно и то же значение a. Тогда $\mathrm{M}(X) = a$ и, значит, $\mathrm{Cov}(X,Y) = \mathrm{M}(XY) - \mathrm{M}(X)\cdot\mathrm{M}(Y) = \mathrm{M}(aY) - a\,\mathrm{M}(Y) = 0$, так что (3.14) примет вид: $0 \le 0$, т.е. всё равно истинно. □

Следующая теорема дополняет теорему 3.3; она указывает, когда в неравенстве (3.16) для коэффициента корреляции достигаются крайние границы. Поскольку в этом неравенстве идёт речь о коэффициенте корреляции между переменными X и Y, мы предполагаем, что дисперсии $\mathrm{D}(X)$ и $\mathrm{D}(Y)$ отличны от нуля.

Теорема 3.4. *Коэффициент корреляции между переменными X и Y равен 1 или* -1 *тогда и только тогда, когда существуют такие числа* $k \ne 0$ *и b, что* $Y = kX + b$, *т.е. между переменными существует линейная зависимость. При этом знак коэффициента корреляции совпадает со знаком коэффициента k.*

Доказательство. Допустим, что $Y = kX + b$, $k \ne 0$. Тогда $\mathrm{M}(Y) = k\,\mathrm{M}(X) + b$, $\mathrm{D}(Y) = k^2\,\mathrm{D}(X)$ и, значит,

$$\begin{aligned}\mathrm{Cov}(X,Y) &\equiv \mathrm{M}\Big(\big[X-\mathrm{M}(X)\big]\cdot\big[Y-\mathrm{M}(Y)\big]\Big)\\ &= \mathrm{M}\Big(\big[X-M(X)\big]\cdot\big[kX+b-kM(X)-b\big]\Big)\\ &= \mathrm{M}\Big(k\big[X-M(X)\big]^2\Big) = k\,\mathrm{M}\Big(\big[X-M(X)\big]^2\Big) = k\,\mathrm{D}(X).\end{aligned}$$

Соответственно,

$$\rho(X,Y) \equiv \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{D}(X)\cdot\mathrm{D}(Y)}} = \frac{kD(X)}{\sqrt{\mathrm{D}(X)\cdot k^2\,\mathrm{D}(X)}} = \frac{k}{|k|} \equiv \mathrm{sgn}(k),$$

где $\mathrm{sgn}(k)$ равен $+1$ или -1 в соответствии с тем, положительно или нет число k.

Если $\rho(X;Y) = \pm 1$, то дискриминант квадратного трёхчлена, введённого при доказательстве теоремы 3.3, равен 0. Значит, этот трёхчлен в некоторой точке t_0 обращается в 0: $\mathrm{D}(t_0X+Y) = 0$. Тогда переменная t_0X+Y для всех объектов анализируемой совокупности принимает одно и то же значение b: $t_0X+Y = b \Leftrightarrow Y = kX+b$ (в качестве k мы взяли число $-t_0$). Коэффициент k отличен от 0, т.к. в противном случае переменная Y тождественно бы равнялась b и, значит, имела бы нулевую дисперсию, что невозможно. Применяя первую часть доказательства, можно утверждать, что знак коэффициента корреляции совпадает со знаком коэффициента k. □

4 Смесь числовых наборов

4.1 Введение

Начнём со следующего простого примера. Предположим, у нас есть два числовых набора: [8,11,12] из трёх чисел и [8,9,5,4] из четырёх чисел. Объединим эти наборы в один набор [8,11,12,8,9,5,4] из 7 чисел. Полученный набор мы назовём *смесью* наборов [8,11,12] и [8,9,5,4].

На практике операция такого рода появляется, например, в следующей ситуации. Рассмотрим две группы школьников. Первая группа (первая генеральная совокупность) состоит из трёх школьников с именами Петя, Саша, Миша. Вторая группа (вторая генеральная совокупность) состоит из четырёх школьниц с именами Надя, Катя, Даша, Маша. Для каждой группы были собраны данные о числе решённых ими задач с кратким ответом из первой и второй частей ЕГЭ 2017 г. Предположим, что Петя, Саша, Миша решили 8, 11, 12 задач соответственно (что описывается набором [8, 11, 12]), а Надя, Катя, Даша, Маша – 8, 9, 5, 4 задач соответственно (что описывается набором [8, 9, 5, 4]).

Объединим эти две группы школьников в одну группу. Новая генеральная совокупность состоит из семи школьников. Для каждого школьника из этой объединённой группы мы знаем, сколько задач с кратким ответом из первой и второй частей ЕГЭ 2017 г. решил этот школьник. Для этого нужно просто скопировать соответствующее число из исходных данных. Результат можно представить, например, в табличной форме (Таблица 4.1). Эта таблица означает, что объединённая генеральная совокупность характеризуется набором [8,11,12,8,9,5,4], который является смесью исходных наборов [8,11,12] и [8,9,5,4].

Таблица 4.1

Имя	Петя	Саша	Миша	Надя	Катя	Даша	Маша
Число задач	8	11	12	8	9	5	4

В рассмотренном доля школьников из первой группы равна $3/7$ – это число мы назовём *весом* первой группы и обозначим w_1. Доля школьников из второй группы равна $4/7$ – это вес w_2 торой группы.

В общем случае *формально* смесь наборов определяется следующим образом.

Определение 4.1. Если мы имеем два набора: набор X из n чисел $x_1,\ldots,x_n$ и набор Y из m чисел $y_1,\ldots,y_m$, то их *смесью* называется набор $Z = X \oplus Y = \left[x_1,\ldots,x_n,y_1,\ldots,y_m\right]$ из $n+m$ чисел.

Важно, тем не менее, помнить, что в реальных ситуациях в смеси $X \oplus Y$ наборы $X = \left[x_1,\ldots,x_n\right]$ и $Y = \left[y_1,\ldots,y_m\right]$ являются наборами значений переменных *X* и *Y*, характеризующих *одно и то же свойство* двух разных совокупностей, состоящих из *однотипных* переменных (как в примере, с которого мы начали этот раздел). Если, модифицировать этот пример и предположить, что набор [8, 11, 12] показывает число задач с кратким ответом, решённых Петей, Сашей, Мишей на ЕГЭ, а набор [8, 9, 5, 4] – число этажей в четырёх соседних домах, то формально можно образовать смесь этих наборов, но набор $X \oplus Y$ не будет иметь никакого содержательного смысла.

В смеси $X \oplus Y$ доля чисел из набора *X* равна $\frac{n}{n+m}$ – это число мы назовём *весом* первого набора и обозначим w_1. Доля чисел из набора *Y* равна $w_2 = \frac{m}{n+m}$ – это вес второго набора. Веса обладают двумя характеристическими свойствами:

- $w_1 \geq 0, w_2 \geq 0$;
- $w_1 + w_2 = 1$.

Смесь наборов можно было бы назвать и «объединением наборов», имея в виду некоторую аналогию с операцией объединения множеств, но это может привести к путанице, т.к. в множестве не может быть повторяющихся элементов, а в наборах дискретных данных повторение – типичная ситуация. Например, смесью наборов [2,3,4] и [4,5] будет набор [2,3,4,4,5] из пяти чисел, в то время как объединение множеств {2,3,4} и {4,5} – это множество {2,3,4,5} из четырёх элементов. Поэтому в статистике приняты термины «смесь наборов», «смешивание наборов», «смесь распределений». Для набора $Z = \left[x_1,\ldots,x_n,y_1,\ldots,y_m\right]$ напрашивается и термин «сумма наборов», но он употребляется только для результата суммирования соответствующих значений двумерной переменной (как мы определили в предыдущем разделе). Поскольку символ «+» уже зарезервирован для обозначения суммы $X + Y$ наборов *X* и *Y*, для обозначения результата смешивания наборов мы используем необычный символ $\oplus$. Отметим, что хотя это обозначение и не является общеупотребительным, оно очень хорошо соответствует сути дела.

Интересно отметить, что для квадрата смеси числовых наборов верно равенство $\left(X \oplus Y\right)^2 = X^2 \oplus Y^2$, в то время как для квадрата суммы наборов (являющихся значениями координат некоторой двумерной переменной) верно привычное по внешней форме равенство $\left(X + Y\right)^2 = X^2 + 2XY + Y^2$.

Действительно, набор $Z = X \oplus Y$ имеет вид $\left[x_1,\ldots,x_n,y_1,\ldots,y_m\right]$. Квадрат набора *Z* (по определению) – это набор Z^2, образованный квадратами чисел z_i из набора *Z*. Поэтому набор Z^2 имеет вид: $\left[x_1^2,\ldots,x_n^2,y_1^2,\ldots,y_m^2\right]$. С другой стороны, набор X^2 имеет вид: $\left[x_1^2,\ldots,x_n^2\right]$, а набор Y^2 – $\left[y_1^2,\ldots,y_m^2\right]$. Поэтому их смесь $X^2 \oplus Y^2$ – это набор $\left[x_1^2,\ldots,x_n^2,y_1^2,\ldots,y_m^2\right]$, т.е. тот же набор, который даёт формула $\left(X \oplus Y\right)^2$.

4.2 Статистические характеристики смеси наборов

Поскольку смесь наборов определяется исходными наборами, естественно ожидать, что и статистические характеристики смеси выражаются через статистические характеристики исходных наборов. Для среднего значения и дисперсии это действительно так. Об этом говорят следующие две теоремы.

Теорема 4.1 (о среднем смеси). *Среднее значение смеси $X \oplus Y$ равно взвешенной сумме средних значений компонент X и Y:*

$$\mathrm{M}(X \oplus Y) = w_1 \cdot \mathrm{M}(X) + w_2 \cdot \mathrm{M}(Y), \qquad (4.1)$$

где веса $w_1 = \dfrac{n}{n+m}$ и $w_2 = \dfrac{m}{n+m}$ пропорциональны размерам n и m этих компонент.

Доказательство. Используя определение среднего значения, мы имеем:

$$\mathrm{M}(X \oplus Y) = \frac{x_1 + \ldots + x_n + y_1 + \ldots + y_m}{n+m} = \frac{x_1 + \ldots + x_n}{n+m} + \frac{y_1 + \ldots + y_m}{n+m}$$
$$= \frac{n}{n+m}\frac{x_1 + \ldots + x_n}{n} + \frac{m}{n+m}\frac{y_1 + \ldots + y_m}{m} = w_1 \cdot \mathrm{M}(X) + w_2 \cdot \mathrm{M}(Y).$$

□

Чтобы проиллюстрировать применение этой теоремы, рассмотрим следующую задачу.

Задача. Специалист страховой компании подготовил отчёт о результатах работы компании за прошедший день. В отчёте, кроме прочего, говорится, что за день было заявлено 17 страховых случаев, и средний размер ущерба составил 8371 руб. Он уже собирался сдавать отчёт руководителю своего отдела, как ему сообщили о трёх новых страховых случаях, по которым ущерб составил 5430 руб., 9485 руб., 12058 руб. соответственно. Определите новый средний размер ущерба по итогам дня.

Решение. Потери компании по 17 заявленным страховым случаям образуют один набор, X, а потери по трём новым случаям – второй, Y. В задаче требуется определить среднее значение смеси наборов, т.е. в наших обозначениях $\mathrm{M}(X \oplus Y)$. По условию, $\mathrm{M}(X) = 8371$, а $\mathrm{M}(Y)$ легко подсчитать: $\mathrm{M}(Y) = \dfrac{5430 + 9485 + 12058}{3} = \dfrac{26973}{3} = 8991$. Веса наборов таковы: $w_1 = \dfrac{17}{20}$, $w_2 = \dfrac{3}{20}$. Поэтому

$$\mathrm{M}(X \oplus Y) = \frac{17}{20} \cdot 8371 + \frac{3}{20} \cdot 8991 = 8464 \text{ (руб).}$$

Этот же результат можно получить и непосредственно (фактически повторяя рассуждения, использовавшиеся при доказательстве теоремы 1).

Общая сумма ущерба по 17 страховым случаям равна $17 \times 8371 = 142307$ руб. С учётом трёх дороднительно заявленных случаев, общая сумма потерь компании по всем 20 страховым случаям равна

$$142307 + 5430 + 9485 + 12058 = 169280 \text{ руб.}$$

Поэтому новое значение среднего ущерба равно $169280/20 = 8464$ руб.

Теорема 4.2 (о дисперсии смеси). *Дисперсия смеси* $X \oplus Y$ *даётся формулой*:

$$\mathrm{D}(X \oplus Y) = w_1 \cdot \mathrm{D}(X) + w_2 \cdot \mathrm{D}(Y) + w_1 w_2 \left[\mathrm{M}(X) - \mathrm{M}(Y)\right]^2. \qquad (4.2)$$

Доказательство. Пусть $Z = X \oplus Y$. Мы знаем, что $\mathrm{D}(Z) = \mathrm{M}\left(Z^2\right) - \left(\mathrm{M}(Z)\right)^2$, где под набором Z^2 понимается набор, полученный возведением в квадрат всех чисел набора *Z*. Из Теоремы 4.1 мы уже знаем, как M(*Z*) связано с M(*X*) и M(*Y*): $\mathrm{M}(X \oplus Y) = w_1 \cdot \mathrm{M}(X) + w_2 \cdot \mathrm{M}(Y)$. Поскольку квадрат смеси, $Z^2 \equiv \left(X \oplus Y\right)^2$, равен $X^2 \oplus Y^2$, т.е. сам может рассматриваться как смесь наборов (X^2 и Y^2), к нему также можно применить Теорему 4.1:

$$M\left(Z^2\right) = w_1 \cdot M\left(X^2\right) + w_2 \cdot M\left(Y^2\right).$$

Поэтому $\mathrm{D}(Z) = w_1 \cdot \mathrm{M}\left(X^2\right) + w_2 \cdot \mathrm{M}\left(Y^2\right) - \left[w_1 \cdot \mathrm{M}\left(X\right) + w_2 \cdot \mathrm{M}\left(Y\right)\right]^2$. Так как $\mathrm{M}\left(X^2\right) = \mathrm{D}(X) + \left(\mathrm{M}(X)\right)^2$, $\mathrm{M}\left(Y^2\right) = \mathrm{D}(Y) + \left(\mathrm{M}(Y)\right)^2$, после перегруппировки слагаемых мы получим:

$$\begin{aligned}\mathrm{D}(Z) &= w_1 \mathrm{D}(X) + w_2 \mathrm{D}(Y) \\ &+ w_1\left(1 - w_1\right)\left(\mathrm{M}(X)\right)^2 - 2w_1 w_2 \mathrm{M}(X)\mathrm{M}(Y) + w_2\left(1 - w_2\right)\left(\mathrm{M}(Y)\right)^2.\end{aligned}$$

Используя соотношения $1 - w_1 = w_2$, $1 - w_2 = w_1$, мы окончательно имеем:

$$\begin{aligned}\mathrm{D}(Z) &= w_1 \mathrm{D}(X) + w_2 \mathrm{D}(Y) + w_1 w_2 \left[\left(\mathrm{M}(X)\right)^2 - 2\mathrm{M}(X)\mathrm{M}(Y) + \left(\mathrm{M}(Y)\right)^2\right] \\ &= w_1 \mathrm{D}(X) + w_2 \mathrm{D}(Y) + w_1 w_2 \left[\mathrm{M}(X) - \mathrm{M}(Y)\right]^2.\end{aligned}$$

□

Замечание 1. Последнее слагаемое в правой части (4.2) можно преобразовать к виду:

$$w_1\left[\mathrm{M}(X \oplus Y) - \mathrm{M}(X)\right]^2 + w_2\left[\mathrm{M}(X \oplus Y) - \mathrm{M}(Y)\right]^2.$$

Действительно, в силу (4.1) это выражение равно

$$w_1\left[w_1 \mathrm{M}(X) + w_2 \mathrm{M}(Y) - \mathrm{M}(X)\right]^2 + w_2\left[w_1 \mathrm{M}(X) + w_2 \mathrm{M}(Y) - \mathrm{M}(Y)\right]^2.$$

Приводя подобные слагаемые в квадратных скобках, мы получим:

$$w_1\left[w_2 \mathrm{M}(Y) - (1 - w_1)\mathrm{M}(X)\right]^2 + w_2\left[w_1 \mathrm{M}(X) - (1 - w_2)\mathrm{M}(Y)\right]^2.$$

Поскольку $1 - w_1 = w_2$, $1 - w_2 = w_1$, мы окончательно имеем:

$$w_1\left[w_2\,\mathrm{M}(Y)-w_2\,\mathrm{M}(X)\right]^2+w_2\left[w_1\,\mathrm{M}(X)-w_1\,\mathrm{M}(Y)\right]^2$$
$$=w_1w_2^2\left[\mathrm{M}(Y)-\mathrm{M}(X)\right]^2+w_1^2w_2\left[\mathrm{M}(X)-\mathrm{M}(Y)\right]^2$$
$$=w_1w_2(w_1+w_2)\left[\mathrm{M}(X)-\mathrm{M}(Y)\right]^2=w_1w_2\left[\mathrm{M}(X)-\mathrm{M}(Y)\right]^2.$$

Замечание 2. В формуле (4.2) сумма двух первых слагаемых, $w_1\cdot\mathrm{D}(X)+w_2\cdot\mathrm{D}(Y)$, задаёт некоторое число, расположенное между $\mathrm{D}(X)$ и $\mathrm{D}(Y)$ (взвешенное среднее дисперсий). Если дисперсии исходных наборов не очень большие, небольшим будет и это среднее. Однако последнее слагаемое, $w_1w_2\left[\mathrm{M}(X)-\mathrm{M}(Y)\right]^2$, может быть очень большим, если средние значения исходных наборов сильно различаются, т.е. наборы X и Y описывают разнородные генеральные совокупности. Иначе говоря, если в ходе статистического анализа некоторой переменной Z выясняется, что её дисперсия велика, то часто это означает, что генеральная совокупность, объекты которой характеризует эта переменная, является объединением двух групп с разными статистическими свойствами. Каждая группа состоят из более-менее однородных объектов, так что в пределах каждой группы разброс значений вокруг своего среднего относительно невелик, но сами средние сильно различаются.

4.3 Распределение значений смеси наборов

Рассмотрим дискретные числовые переменные X и Y, характеризующие *одно и то же свойство* двух разных совокупностей, состоящих из n и m *однотипных* объектов соответственно. Пусть $u_1,\ldots,u_N$ – теоретически возможные значения переменных X и Y, а $f_X(u_1),\ldots,f_X(u_N)$ и $f_Y(u_1),\ldots,f_Y(u_N)$ – распределения этих переменных. Таким образом, набор $\left[x_1,\ldots,x_n\right]$ состоит из

$$t_X(u_1)=f_X(u_1)\cdot n \text{ чисел } u_1,\ \ldots,\ t_X(u_N)=f_X(u_N)\cdot n \text{ чисел } u_N,$$

а набор $\left[y_1,\ldots,y_m\right]$ – из

$$t_Y(u_1)=f_Y(u_1)\cdot n \text{ чисел } u_1,\ \ldots,\ t_Y(u_N)=f_Y(u_N)\cdot n \text{ чисел } u_N.$$

Как мы отмечали, некоторые из чисел $t_X(u)$ и $t_Y(u)$ могут быть нулевыми (переменные X, Y не обязаны принимать все теретически возможные значения).

Очевидно, что набор $Z=X\oplus Y$ состоит только из чисел из списка $u_1,\ldots,u_N$, причём каждое значение u в наборе Z появляется $t_X(u)+t_Y(u)$ раз:

$$t_Z(u)=t_X(u)+t_Y(u).$$

Разделим это равенство на число $n+m$ объектов в наборе $Z=X\oplus Y$ и проведём преобразования, аналогичные тем, которые мы проделали при доказательстве Теоремы 4.1. Поскольку $\dfrac{t_Z(u)}{n+m}=f_Z(u)$, мы получим:

$$f_Z(u) = \frac{t_X(u) + t_Y(u)}{n+m} = \frac{n}{n+m}\frac{t_X(u)}{n} + \frac{m}{n+m}\frac{t_Y(u)}{m} = w_1 f_X(u) + w_2 f_Y(u).$$

Итак, мы доказали следующую теорему.

Теорема 4.3. *Распределение смеси $X \oplus Y$ наборов X и Y может быть найдено по распределениям исходных наборов с помощью формулы*

$$f_{X\oplus Y}(u) = w_1 f_X(u) + w_2 f_Y(u). \tag{4.3}$$

Операция смешивания наборов может кардинально изменить характер распределения исходных наборов. Рассмотрим, например, следующие наборы X (из $n = 25$ чисел) и Y (из $m = 50$ чисел):

$$X = \left[1,2,2,2,2,\underbrace{3,\ldots,3}_{12\text{ чисел}},4,4,4,4,5,5,6,7\right],$$

$$Y = \left[2,3,3,4,4,4,4,\underbrace{5,\ldots,5}_{8\text{ чисел}},\underbrace{6,\ldots,6}_{24\text{ числа}},\underbrace{7,\ldots,7}_{8\text{ чисел}},8,8,9\right].$$

Их распределения приведены во второй и третьей строках Таблицы 4.2 соответственно (мы считаем, что возможными значениями являются натуральные числа от 1 до 9). В последней строке Таблицы 4.2 приведено распределение смеси этих наборов (числа округлены до сотых долей).

Таблица 4.2

u	1	2	3	4	5	6	7	8	9
$f_X(u)$	1/25	4/25	12/25	4/25	2/25	1/25	1/25	0	0
$f_Y(u)$	0	1/50	2/50	4/50	8/50	24/50	8/50	2/50	1/50
$f_{X\oplus Y}(u)$	1/75	5/75	14/75	8/75	10/75	25/75	9/75	2/75	1/75

Столбиковая диаграмма на Рисунке 4.1 изображает распределение значений набора X, столбиковая диаграмма на Рисунке 4.2 – распределение значений набора Y, а столбиковая диаграмма на Рисунке 4.3 – распределение значений набора $X \oplus Y$.

Хорошо видно, что исходные наборы X и Y имеют ярко выраженную единственную моду, т.е. являются *унимодальными*. Однако их смесь имеет две моды. Такое распределение называется *бимодальным*. Обычно бимодальность распределения свидетельствует о том, что это распределение является смесью унимодальных. Иначе говоря, генеральная совокупность, объекты которой характеризует эта переменная, является объединением двух групп с разными статистическими свойствами. Каждая группа состоят из более-менее однородных объектов, так что в пределах каждой группы разброс значений вокруг своего среднего относительно невелик, но сами средние сильно различаются (сравните это замечание с Замечанием 2 к конце Раздела 4.2).

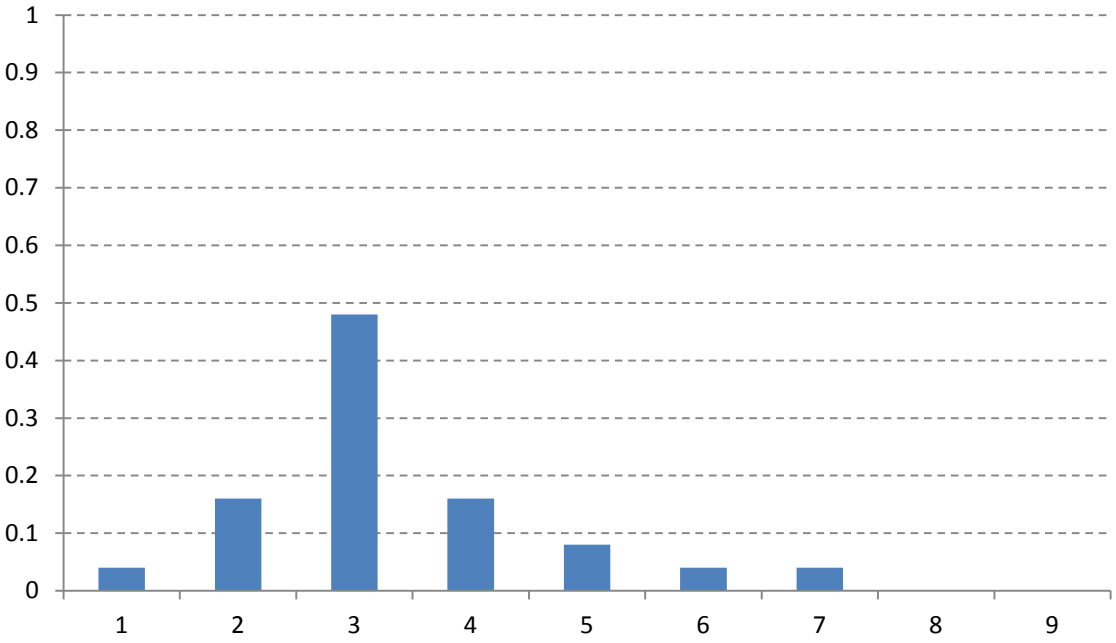

Рисунок 4.1 Распределение $f_X(u)$

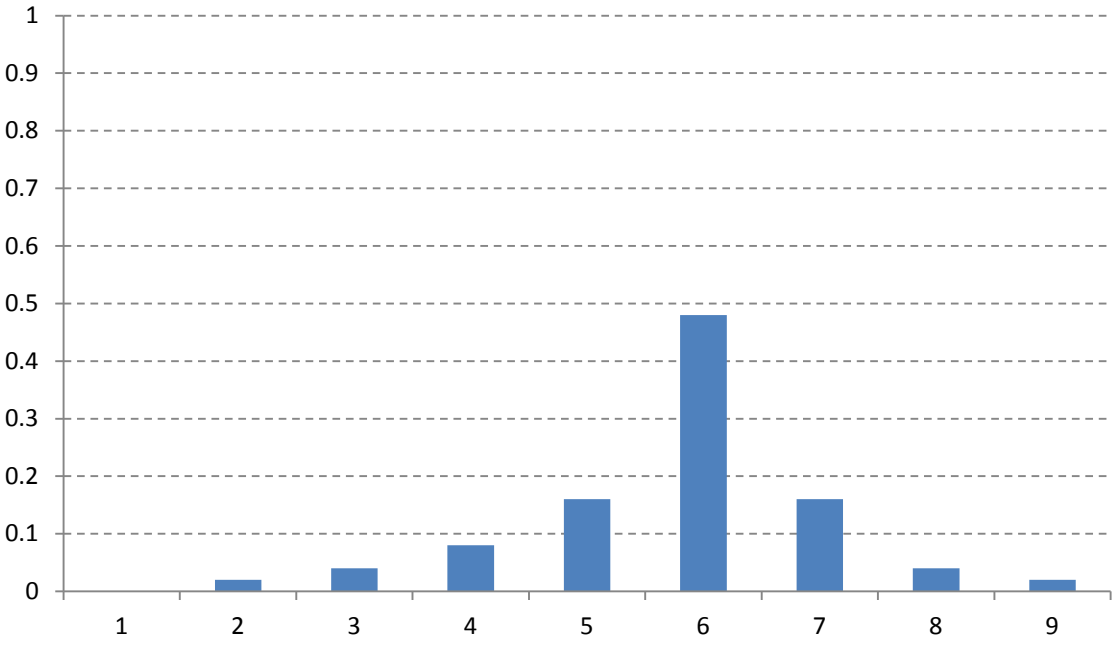

Рисунок 4.2 Распределение $f_Y(u)$

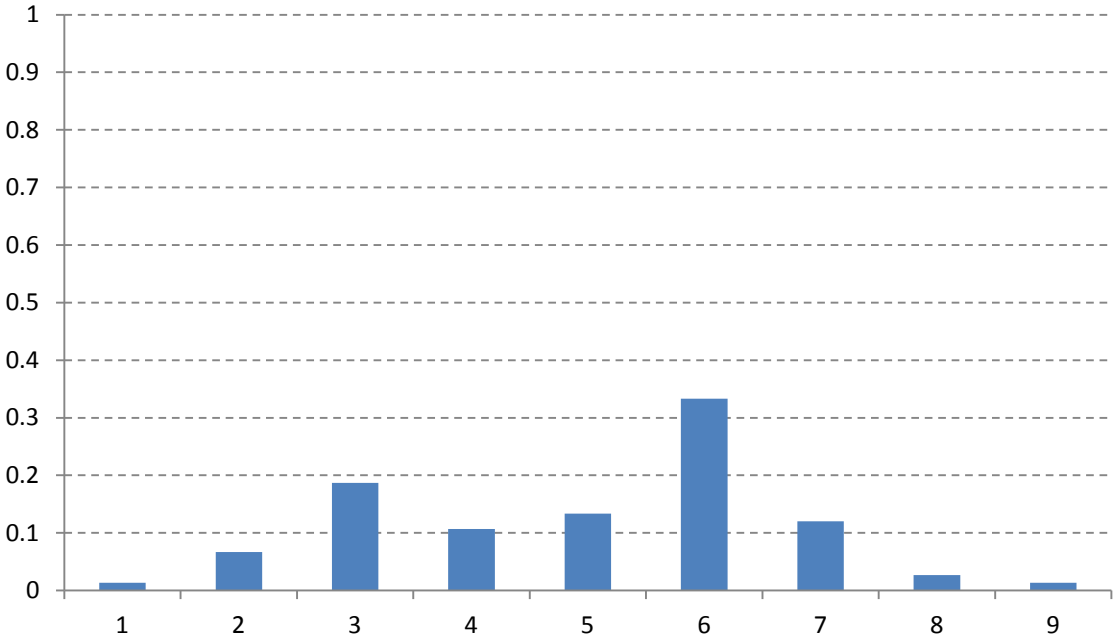

Рисунок 4.3 Распределение $f_{X\oplus Y}(u)$

5 Заключение. Роль в статистики в естествознании

5.1 Статистика и открытие аргона

5.1.1 Введение. Людей всегда интересовал состав воздуха, которым они дышат, но только великий французский учёный Антуан-Лоран Лавуазье установил (примерно в 1775 г.), что атмосферный воздух является смесью двух газов; один из них пригоден для дыхания и поддерживает горение, а второй – нет. Первый газ Лавуазье назвал кислородом, а второй – азотом. Следует отметить, что эти элементы были открыты за несколько лет до Лавуазье: кислород – шведским химиком Шееле (разложением оксида азота) и независимо от него английским химиком Пристли (разложением оксида ртути), азот – британским учёным Кавендишем (химическими опытами с воздухом). Кроме того, Лавуазье отметил, что атмосферный воздух должен содержать и некоторое количество других газов и веществ, которые могут испаряться и находиться в газообразном состоянии при обычных температуре и давлении (в качестве примера можно упомянуть углекислый газ и воду).

Больше ста лет учёные были твёрдо уверены, что за исключением водяных паров, углекислого газа, озона и т.п. газообразных веществ, содержание которых крайне незначительно, атмосферный воздух является смесью двух газов: азота (примерно 79% по объёму) и кислорода (примерно 21% по объёму). Например, в знаменитой книге «Основы химии» Д.И.Менделеева (первое издание, Спб, 1869-1871) в главе 11 «Азот и воздух» мы читаем:

«**Об атмосферном воздухе.** Атмосферный воздух содержит смесь нескольких газов и парообразных веществ. Одни из них встречаются в нем почти всегда в одинаковых пропорциях, другие же, напротив того, очень изменчивы в своем содержании. Главнейшие составные части воздуха, встречающиеся в нем постоянно и исчисленные в последовательном порядке своего относительного количества, суть следующие: азот, кислород, водяной пар, угольная кислота, азотная кислота, пары аммиачных солей… во 100 частях воздуха по объему заключается кислорода 20.85, а азота 79.15.»

Сумма 20.85+79.15 даёт 100, так что Менделеев фактически подчёркивал пренебрежимо малое сдержание в воздухе других перечисленных им газов. На самом деле, как мы сейчас знаем, почти 1% (точнее, 0.934%) объёма воздуха занимает аргон. Суммарное объёмное содержание других газов в чистом сухом воздухе крайне мало. Четвёртой (в порядке убывания объёмного содержания) составной частью воздуха является углекислый газ. Хотя его концентрация не является постоянной, в среднем в чистом сухом воздухе она равна примерно 0.03% (0.04% по самым последним данным; рост связан с деятельностью человека). Суммарное объёмное содержание остальных газов – несколько тысячных долей процента.

Плотность аргона при нормальных условиях равна примерно 1.784 кг/куб.м. Поэтому, например, в квартире площадью 70-75 кв.м. (объём около 200 куб.м.) содержится почти 3.5 кг аргона. Совершенно поразительно, что

учёные долго не могли обнаружить в воздухе газ, масса которого в обычном помещении измеряется килограммами.

5.1.2 Опыты лорда Рэйли. Открытие этой третьей значительной составной части воздуха началось в 1882 году, когда английский физик лорд Рэйли (при рождении Джон Вильям Страт, с 1873 г. – 3-й барон Рэйли) решил провести опыты по уточнению плотностей воздуха, кислорода и азота. Методику проведения исследования и основные результаты он опубликовал в нескольких научных работах. Для нас самыми важными представляются две статьи:

1. Lord Rayleigh. On the Densities of the Principal Gases. *Proceedings of the Royal Society of London*, 1893, vol. 53, pp. 134-149;
2. Lord Rayleigh. On an Anomaly Encountered in Determinations of the Density of Nitrogen Gas. *Proceedings of the Royal Society of London*, 1894, Vol. 55, pp. 340-344,

в которых приведены результаты опытов и их статистический анализ.

Основным элементом лабораторной установки лорда Рэйли был сферический сосуд объёмом 1.83652 л; объём он измерил взвешивая пустой сосуд и сосуд, заполненный водой. Этот сосуд он заполнял изучаемым газом, а затем при температуре 0°C и нормальном давлении газа в сосуде измерял массу сосуда с газом. Вычитая массу пустого сосуда, он находил массу газа в сосуде.

Для каждого газа (воздух, кислород, «азот») лорд Рэйли проводил серию опытов, что давало набор из нескольких чисел.

Опыты по определению плотности воздуха. Первая группа опытов лорда Рэйли была посвящена определению плотности сухого чистого воздуха (из сельской местности), из которого удалён углекислый газ; для этого воздух пропускали через раствор поташа (карбоната калия K_2CO_3). Измерения проводились при температуре 0° и нормальном давлении. Полученные результаты приведены в Таблице 5.1.

Таблица 5.1 Измерение плотности воздуха

№ п/п, i	Дата проведения опыта	Масса воздуха в i-м опыте, m_i^{air} (г)	$y_i^{\text{air}} = (m_i^{\text{air}} - 2.376)\cdot 10^4$
1	27.09.1892	2.37686	8.6
2	29.09.1892	2.37651	5.1
3	03.10.1892	2.37653	5.3
4	08.10.1892	2.37646	4.6
5	11.10.1892	2.37668	6.8
6	13.10.1892	2.37679	7.9
7	15.10.1892	2.37647	4.7

Эти опыты были абсолютно идентичны по методике проведения и лорд Рэйли чрезвычайно тщательно следил за тем, чтобы они копировали друг друга в мельчайших деталях. Однако, как обычно, неизбежные случайные погрешности

привели к определённым отличиям в результатах. Поэтому лорд Рэйли провёл (используя современный язык) небольшой статистический анализ полученных данных. Прежде всего, чтобы сгладить случайные колебания, он вычислил среднее значение:

$$\mathrm{M} = \frac{m_1^{\mathrm{air}} + \ldots + m_7^{\mathrm{air}}}{7} = 2.37661 \text{ (г)}.$$

Для того, чтобы понять, насколько значительно отличаются результаты разных опытов друг от друга, он вычислил размах:

$$\mathrm{R} = \max\left(m_i^{\mathrm{air}}\right) - \min\left(m_i^{\mathrm{air}}\right) = 2.37686 - 2.37646 = 0.00040 \text{ (г)} = 0.4 \text{ (мг)}$$

и сравнил его со средним значением: $\mathrm{R}/\mathrm{M} \approx 1/6000$. Столь малое *относительное* значение разброса чисел анализируемого набора вокруг среднего позволило ему сделать вывод о том, что вес воздуха в сосуде измерен достаточно точно. Кроме того, дробь 1/6000 говорит, что на счёт неизбежных погрешностей при проведении экспериментов можно отнести только отличия такого порядка; бóльшие отличия должны иметь другие объяснения.

Если в качестве значения массы воздуха в сосуде взять среднее значение 2.37661, то плотность воздуха равна

$$\frac{2.37661 \text{ (г)}}{1.83652 \text{ (л)}} \approx 1.294 \text{ г/л},$$

что практически совпадает с общепринятым сейчас значением. На самом деле этот последний этап расчётов должен быть сложнее, т.к. лорд Рэйли учитывал определённые, точно контролируемые, нюансы в процессе изменения величин. Впрочем, это практически не влияет на окончательное значение плотности воздуха.

Изложенная нами небольшая статистическая теория позволяет нагляднее изложить рассуждения лорда Рэйли. Беглый взгляд на таблицу 5.1 показывает, что все числа m_i^{air} расположены возле числа 2.376. Вычитая из чисел m_i^{air} число 2.376 мы получим набор 0.00086, 0.00051, … из очень маленьких чисел. Применим поэтому следующее преобразование: $y_i^{\mathrm{air}} = (m_i^{\mathrm{air}} - 2.376) \cdot 10^4$ (множитель 10^4 превращает десятичные дроби с 5 знаками после запятой в числа умеренной величины). Значения переменной Y^{air} приведены в последнем столбце таблицы 5.1. Среднее значение новой переменной $\mathrm{M}\left(Y^{\mathrm{air}}\right) = 6.1$, а размах $\mathrm{R}\left(Y^{\mathrm{air}}\right) = 8.6 - 4.6 = 4$. Для наглядности на Рисунке 5.1 мы изобразили значения переменной Y^{air} точками на числовой оси и там же показали её среднее значение и размах.

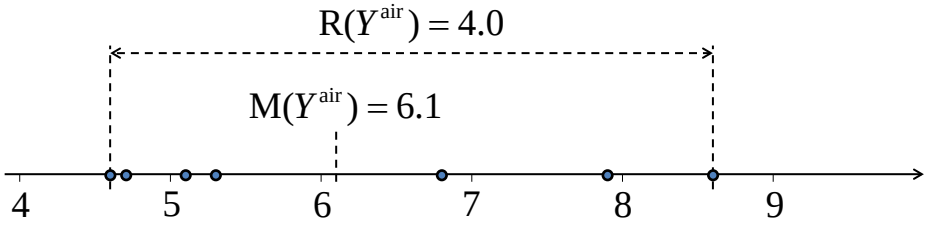

Рисунок 5.1

Из этого рисунка хорошо виден характер разброса значений переменной Y^{air} вокруг её среднего значения из-за случайных погрешностей при проведении разных экспериментов. Напомним, что в абсолютных цифрах точки на рисунке изображают десятитысячные доли грамма, т.е. десятые доли миллиграмма, в то время как основная переменная (масса воздуха в сосуде) равна примерно 2 грамма.

Опыты по определению плотности кислорода. Примерно такой же статистический анализ провёл лорд Рэйли и после серии опытов по измерению массы сосуда, наполненного кислородом (всего 16 опытов), но мы не будем на них останавливаться, а перейдём к самой интересной части его исследований – опытам по определению плотности азота.

Опыты по определению плотности азота. Для получения азота лорд Рэйли использовал два принципиально разных метода.

В первом методе источником азота был сухой чистый воздух (из сельской местности). Из этого воздуха удаляли углекислый газ (пропуская воздух через раствор поташа) и кислород (пропуская воздух через раскалённую медь, железо и т.п.). После этого, по воззрениям того времени, получался чистый азот. На самом деле, как мы сейчас знаем, получалась смесь азота и аргона (и крайне незначительного количества других газов). Именно по этой причине мы и поставили выше слово «азот» в кавычки. Кислород из воздуха удалялся разными методами; мы условно обозначим их метод A, B, C, D.

В Таблице 5.2 приведены результаты опытов, по получению «атмосферного азота» которые лорд Рэйли провёл в 1892 – 1894 гг.

Таблица 5.2
Вес сосуда с «азотом» (источник «азота» – атмосферный воздух)

№ п/п, i	Дата проведения опыта	Метод удаления кислорода	Масса «азота», x'_i (г)	y'_i
1	08.08.1892	A	2.31035	3.5
2	10.08.1892	A	2.31026	2.6
3	15.08.1892	A	2.31024	2.4
4	17.09.1892	B	2.31012	1.2

5	20.09.1892	B	2.31027	2.7
6	12.12.1893	C	2.31017	1.7
7	14.12.1893	C	2.30986	-1.4
8	19.12.1893	C	2.31010	1
9	22.12.1893	C	2.31001	0.1
10	01.01.1894	D	2.31163	16.3
11	04.01.1894	D	2.30956	-4.4
12	27.01.1894	E	2.31024	2.4
13	30.01.1894	E	2.31010	1
14	01.02.1894	E	2.31028	2.8

Поскольку кислород удалялся разными методами, набор из 14 чисел в четвёртом столбце Таблицы 5.2 на самом деле является смесью пяти наборов. Чтобы понять, одинаковы ли статистические характеристики этих наборов, лорд Рэйли вычислил среднее значение для каждого метода удаления кислорода: $\mathrm{M}^A = 2.31028$, $\mathrm{M}^B = 2.31020$, $\mathrm{M}^C = 2.31003$, $\mathrm{M}^D = 2.31059$, $\mathrm{M}^E = 2.31020$. Различаются эти средние только в четвёртом знаке после десятичной точки (разница между наибольшим и наименьшим средним составляет 0.55мг), что примерно соответствует размаху колебаний внутри каждого набора (для метода A размах равен 0.11мг, B – 0.15мг, C – 0.31мг, D – 2.07мг, E – 0.18мг). Поэтому разные способы удаления кислорода из воздуха дают примерно один и тот же результат, так что объединение пяти наборов в один вполне обосновано.

Изложенная нами небольшая статистическая теория позволяет нагляднее изложить рассуждения лорда Рэйли. Беглый взгляд на таблицу 5.2 показывает, что все числа x_i' расположены возле числа 2.31. Вычитая из чисел x_i' число 2.31 мы получим набор 0.00035, 0.00026, … из очень маленьких чисел. Применим поэтому следующее преобразование: $y_i' = (x_i' - 2.31)\cdot 10^4$ (множитель 10^4 превращает десятичные дроби с 5 знаками после запятой в числа умеренной величины). Переменная Y' показывает абсолютное отклонение результата опыта от числа 2.31(грамма) в десятых долях миллиграмма. Значения этой переменной приведены в последнем столбце таблицы 5.2. Для наглядности на Рисунке 5.2 мы изобразили значения переменной Y'' в виде точечной диаграммы (на оси абсцисс указан номер опыта). Из этого рисунка хорошо видно, что группы точек, соответствующие разным сериям опытов, можно рассматривать как часть общей картины. Из этой картины выбивается только опыт №10 (от 1 января 1894 г.), что лорд Рэйли объясняет невозможностью провести точное измерение в этот день. Иначе говоря, объединение пяти наборов в один вполне обосновано.

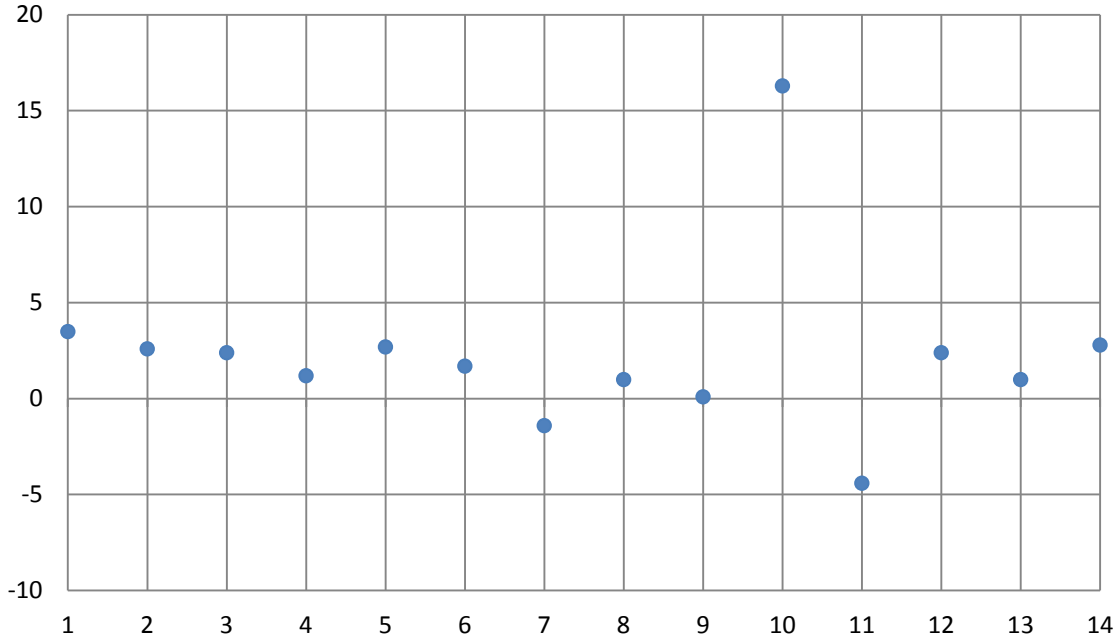

Рисунок 5.2

Во втором методе источником азота были различные химические соединения, содержащие азот: оксид азота NO (метод A), закись азота N_2O (метод B), закись азота N_2O + через выделенный азот дополнительно пропускался электрический разряд (метод B+), нитрат аммония NH_4NO_3 (метод C), из которых с помощью химических реакций выделялся практически чистый азот. В Таблице 5.3 приведены результаты опытов по получению «химического азота», которые лорд Рэйли провёл в 1892 – 1894 гг. (в первом столбце мы продолжаем нумерацию опытов, начатую в Таблице 5.2).

Таблица 5.3
Вес сосуда с азотом (источник азота – химическое соединение)

№ п/п, i	Дата проведения опыта	Метод: Источник азота	Масса азота, x_i'' (г)	y_i''
15	29.11.1893	A: NO	2.30143	–85.7
16	02.12.1893	A: NO	2.29890	–111
17	05.12.1893	A: NO	2.29816	–118.4
18	06.12.1893	A: NO	2.30182	–81.8
19	26.12.1893	B: N_2O	2.29869	–113.1
20	28.12.1893	B: N_2O	2.29940	–106
21	02.01.1894	B+: N_2O	2.30074	–92.6
22	05.01.1894	B+: N_2O	2.30054	–94.6
23	09.01.1894	C: NH_4NO_3	2.29849	–115.1
24	13.01.1894	C: NH_4NO_3	2.29889	–111.1

Поскольку азот выделялся из разных химических соединений, набор из 10 чисел в четвёртом столбце Таблицы 5.3 на самом деле является смесью четырёх наборов. Чтобы понять, что статистические характеристики этих наборов одинаковы, лорд Рэйли вычислил среднее значение для каждого из них: $\mathrm{M}^{A} = 2.30008$, $\mathrm{M}^{B} = 2.29904$, $\mathrm{M}^{B+} = 2.30064$, $\mathrm{M}^{C} = 2.29869$. Хотя, как отметил лорд Рэйли, эти средние расположены не так близко друг к другу, как аналогичные средние для разных методов получения «атмосферного азота», согласие между ними достаточно хорошее – разница между наибольшим и наименьшим средними значениями меньше 0.002г=2мг, что составляет меньше 0.1% от типичной массы газа в сосуде (эта величина имеет порядок 2.3 г). Поэтому разные способы получения азота из химических соединений дают примерно один и тот же результат, так что объединение четырёх наборов в один вполне обосновано.

Как и в случае «атмосферного азота», изложенная нами небольшая статистическая теория позволяет нагляднее изложить рассуждения лорда Рэйли и подтврдить его вывод. Применяя то же преобразование $y_i'' = (x_i'' - 2.31)\cdot 10^4$, что и в случае «атмосферного азота», мы получим набор из десяти чисел y_i'' (см. последний столбец таблицы 5.3). На Рисунке 5.3 значения переменной Y'' представлены графически в виде точечной диаграммы (на оси абсцисс указан номер опыта). Из этого рисунка хорошо видно, что группы точек, соответствующие разным сериям опытов, можно рассматривать как часть общей картины. Иначе говоря, объединение четырёх наборов в один вполне обосновано.

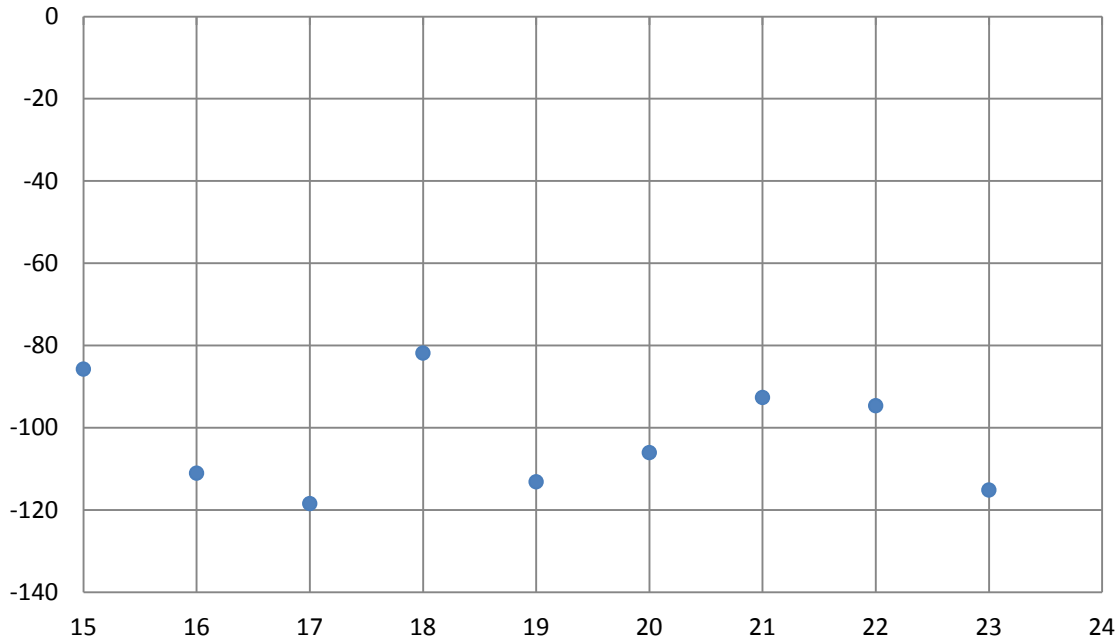

Рисунок 5.3

Беглый взгляд на Таблицы 5.2 и 5.3 показывает, что эти различия в результатах разных опытов несущественны и в качестве массы азота в сосуде можно взять среднее значение по результатам всех 24 опытов; оно равно 2.305844. Для плотности азота это дало бы значение $\frac{2.305844\ (\text{г})}{1.83652\ (\text{л})} \approx 1.25555$ г/л, что практически совпадает с точным значением 1.251 г/л (относительная погрешность меньше 0.4%).

Однако лорд Рэйли не стал сразу смешивать оба набора чисел в один набор, а решил провести дополнительный статистический анализ результатов своих опытов. С этой целью он нашёл средние для каждого набора. Для «азота», полученного из атмосферного воздуха, среднее равно 2.310228 (г), а для азота, полученного из химических соединений, среднее равно 2.299706 (г). Разница между ними составляет 0.010522(г)≈10мг. Сама по себе эта разница очень мала как в абсолютном выражении, так и по отношению к средней массе – отношение примерно равно 1/200. Однако если эту разницу сравнивать с разницей между наибольшим и наименьшим средним для разных методов получения «атмосферного азота» (0.55мг) и аналогичной величиной для разных методов получения «химического азота» (2 мг), то следует признать, что величина 10мг хотя и «мала сама по себе, но лежит вне пределов ошибок эксперимента и может быть объяснена только различием в характере газа». Эта цитата взята из письма, которое лорд Рэйли написал 24 сентября 1892 года редактору журнала Nature (Rayleigh. *Density of Nitrogen. Letter to the Editor*. Nature, 1892, vol. 46, No.1196, p.512-513). В этом кратком (около трети листа А4) письме лорд Рэйли не привёл никаких числовых данных; они были опубликованы немного позже в упомянутых выше статьях.

На языке статистики наблюдение лорда Рэйли означает, что наборы x_i' из Таблицы 5.2 и x_i'' описывают разные генеральные совокупности. Это хорошо видно из Рисунка 5.4, где мы представили графически в виде точечной диаграммы значения переменной $Y = (X - 2.31)\cdot 10^4$ для всех 24 опытов (как и на Рисунках 5.2, 5.3, на оси абсцисс указан номер опыта).

Из Рисунка 5.4 мы ясно видим, что значения преобразованной переменной распадаются на две группы: первые 15 значений, соответствующие «азоту» из атмосферного воздуха, расположены возле числа 0, в то время как последние 10 значений, соответствующие «азоту» из химических соединений, расположены возле числа –100.

Итак, *с помощью простейшего статистического анализа результатов своих опытов* лорд Рэйли

1. сделал абсолютно верный вывод о том, что «атмосферный азот» явно тяжелеее «химического азота», полученного из химических соединений;
2. высказал гениальную догадку, что «атмосферный азот» чем-то отличается от «химического азота», и предложил учёным выяснить причину этого необъяснимого явления.

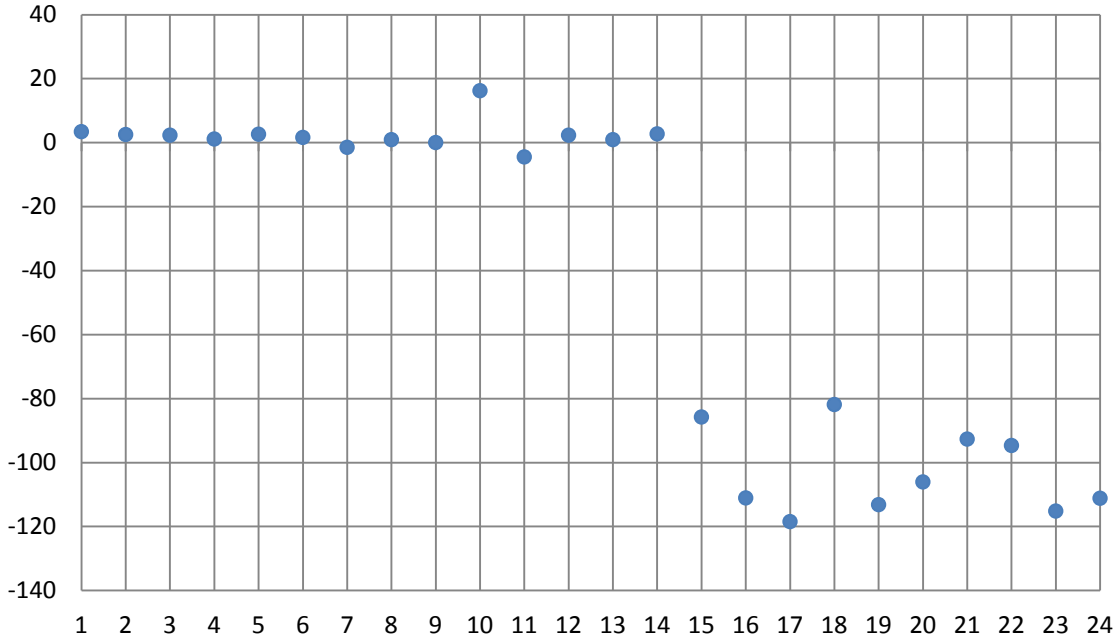

Рисунок 5.4

Лорд Рэйли и присоединившийся к его исследованиям шотландский химик сэр Вильям Рамзи предположили, что это может быть следствием только того, что «азот», полученный из атмосферного воздуха, содержит ещё один, тяжёлый, газ, который не вступает в химические реакции и потому не мог быть удалён применявшимися методами. Дальнейшие исследования, детально описанные в статье Lord Rayleigh and William Ramsay. Argon, a New Constituent of the Atmosphere. Proceedings of the Royal Society of London, Vol. 57, (1894 - 1895), pp. 265-287, быстро показали, что это действительно так, т.е. в состав атмосферного воздуха входит ещё один, до тех пор неизвестный науке, газ, который назвали аргоном. Именно по этой причине мы брали слово «азот» в кавычки – атмосферный «азот» на самом деле был смесью азота и аргона, а также ряда других инертных газов (правда, их содержание было намного меньше).

В 1904 году за «исследования плотностей наиболее важных газов и связанное с этими изысканиями открытие аргона» лорду Рэйли была присуждена Нобелевская Премия по физике, а сэру Рамзаю «в признание его трудов по открытию инертных газов в воздухе и определение их места в периодической системе» – Нобелевская Премия по химии (цитаты взяты из решений Нобелевского Комитета).